Universitat Autònoma de Barcelona

Lyman-alpha Intensity Mapping forecast with the PAU Survey

Pablo RENARD GUIRAL

UNIVERSITAT AUTÒNOMA DE BARCELONA

DOCTORAL THESIS

Lyman-alpha Intensity Mapping forecast with the PAU Survey

Author:
Pablo RENARD GUIRAL

Director:
Enrique GAZTAÑAGA BALBAS
Tutor:
Enrique FERNÁNDEZ SÁNCHEZ

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

Cosmology & Fundamental Physics Department
Institute of Space Sciences (ICE-CSIC/IEEC)

January 19, 2021

UNIVERSITAT AUTÒNOMA DE BARCELONA

Abstract

Department of Physics
Institute of Space Sciences (ICE-CSIC/IEEC)

Doctor of Philosophy

Lyman-alpha Intensity Mapping forecast with the PAU Survey

by Pablo RENARD GUIRAL

Versió catalana

En els pròxims anys, la quantitat de dades observacionals disponibles per l'astrofísica en conjunt, i particularment la cosmologia, hi augmentarà en ordres de magnitud degut als nous programes observacionales i als avançaments tecnològics. Per aprofitar totalment aquestes observacions venideres, nous mètodes estadistics per a obtenir resultats científics han de ser desenvolupats i aplicats. Intensity Mapping és una d'aquests nous mètodes, que consisteix a traçar una línia d'emissió en coordenades angulars i redshift, sense considerar objectes resolts.

Aquesta tesi hi està enfocada a la simulació i avaluació d'Intensity Mapping amb la línea Lyα fent la correlació creuada de dos tipus de surveys òptiques: la survey fotomètrica de banda estreta PAUS, i les surveys espectroscòpiques eBOSS i DESI. La primera hauria de contenir emissió Lyα estesa provinent del mitjà intergalàctic al fons de les imatges, mentres que les segones contenen el bosc Lyα (és a dir, el conjunt de línies d'absorció generades pel mitjà intergalàctic) en el seus espectres de quàsars. Al fer la correlació creuada d'ambdues, hauriem d'obtenir una senyal de l'emissió de Lyα difusa, a pesar dels objectes no correlacionats i el soroll a les imatges de PAUS. Per simular aquesta correlació creuada, una simulació hidrodinàmica dissenyada específicament per a l'estudi de Lyα al mitjà intergalàctic ha sigut emprada per simular tant l'emissió Lyα com l'absorció, mentres que els objectes no correlacionats a les imatges de PAUS han sigut modelats utilitzant un catàleg simulat profund al que s'han ajustat plantilles d'emissió espectral als seus objectes. El soroll instrumental i atmosfèric s'ha afegit a les imatges de PAUS simulades mesurant directament el soroll de conjunts d'imatges apilades i emmascarades, i s'ha introduït a la simulació com una distribució gaussiana. A més a més, un codi de correlació creuada ha estat desenvolupat des de zero, optimitzat i validat dins d'aquesta tesi per calcular les correlacions creuades.

Els resultats es presenten de forma probabilística: per casos diferents (PAUS-eBOSS, PAUS-DESI i dues extensions hipotètiques de PAUS) han sigut calculades 1,000 correlacions creuades amb realitzacions diferents del soroll instrumental i atmosfèric de les imatges de PAUS, a més de les posicions dels quàsars. Amb l'aproximació optimista de soroll instrumental no correlacionat (el que requeriria més treball reduint les dades), fins l'extensió més llarga de PAUS que hem considerat té tan sols una probabilitat de detecció del 15%; si emprem el soroll correlacionat que s'observa a les imatges, totes les probabilitats passen a ser negligibles.

A pesar d'aquests resultats negatius, podem extreure algunes conclusions valuoses. Els diferents tipus de funcions de correlació de dos punts que hem considerat (monopol, paral·lela i perpendicular a la línia de visió) mostren comportaments

complementaris: el monopol i la paral·lela tenen un millor rendiment a escales més grans, mentres que la perpendicular sondeja millor les escales menors de 10 Mpc/h. A més a més, el suavitzat en redshift a la senyal de Ly-alpha degut als filtres fotomètrics ha sigut implementat de forma adecuada al model teòric, i fins i tot pot ser emprada per a predir les escales amb millor ràtio entre soroll i senyal per a cada funció de correlació. Per tant, aquests descobriments indiquen que aquesta metodologia pot ser més adequada per les surveys de banda ampla; concloem la tesi donant algunes directrius bàsiques sobre com aquest estudi es pot replicar per aquest tipus de dades.

English version

In the upcoming years, the amount of observational data available for astrophysics as a whole, and particularly cosmology, is expected to increase by orders of magnitude due to the new observational programs and technological advancements. To fully capitalise on these upcoming observations, new statistical methods to draw scientific results have to be developed and applied. Intensity Mapping is one of such new methods, which consists on the tracing of a sharp spectral feature (e.g., a emission line) in angular coordinates and redshift, without considering resolved objects.

This thesis is focused on the simulation and evaluation of Intensity Mapping with the Lyα line by cross-correlating two kinds of optical surveys: the narrow-band photometric survey PAUS, and the spectroscopic surveys eBOSS and DESI. The former should contain extended Lyα emission from the intergalactic medium in the background of the images, while the latter contains the Lyα forest (i.e., the set of absorption lines generated by the intergalactic medium) in their quasar spectra. Cross-correlating both then should yield a signal of the diffuse Lyα emission, despite the foregrounds and noise in PAUS images.

In order to simulate this, a hydrodynamic simulation specifically designed for the study of Lyα in the intergalactic medium has been used to simulate both Lyα emission and absorption, while the foregrounds in PAUS images have been modeled using a deep lightcone mock catalogue and fitting spectral energy distributions to its objects. Instrumental/atmospheric noise has been added to the simulated PAUS images by directly measuring the noise from sets of masked and stacked images, and introducing it on the simulation as a Gaussian distribution. Besides, a cross-correlation code in Python has been developed from scratch, optimised and validated in the framework of the thesis to compute the cross-correlations.

The results are presented in a probabilistic manner: for different cases (PAUS-eBOSS, PAUS-DESI and two hypothetical PAUS extensions with DESI) 1,000 cross-correlations are computed with different realisations of the instrumental/atmospheric noise of PAUS images, as well as quasar positions. With the optimistic approximation of uncorrelated instrumental noise (which would require further work in data reduction), even the largest PAUS extension considered only yields a probability of detection of $\sim$15%; using the actual correlated noise all the probabilities become negligible.

In spite of these negative results, some valuable conclusions are extracted. The different kinds of two-point correlation functions (monopole, parallel and perpendicular to the line of sight) show complementary behaviours: monopole and parallel perform better at larger scales, while perpendicular samples better the scales smaller than 10 Mpc/h. Besides, the redshift smoothing in the Lyα signal due to

the photometric filters has been properly implemented in the theoretical model, and can even be used to predict the scales with better SNR for each correlation function. Therefore, these findings indicate that this methodology may be more suitable for broad-band surveys; we conclude the thesis giving some basic guidelines on how this study could be replicated for broad-band data.

Acknowledgements

First, I would like to express my deepest gratitude to my project advisor, Enrique Gaztañaga. It has been a pleasure working with him, and all what I have learned under his supervision has been invaluable, both to develop as a cosmologist and, more broadly, as a researcher. It is thanks to him that I can consider myself somewhat of a scientist now. Furthermore, I must thank him also for accepting me as a PhD student even without a grant (and trusting me enough to send me abroad for so long!). Without this opportunity, I believe that a research career in cosmology (which seemed like a pipe dream some years ago) would have been far less likely for me.

On a similar note, I have to thank as well Rupert Croft for his collaboration on the work presented in this thesis, as well as the non-negligible amount of recommendation letters that I asked him. His pioneering work on Lyα IM, as well as his insights on this thesis have been inspiring, and he has also been very friendly and encouraging in all of our interactions.

Besides, from the staff of the Cosmology department at ICE I would also like to thank Francisco Castander and Ricard Casas, mainly for observing with me at the William Herschel Telescope for PAUS, and showing me the basics of photometric observations. In spite of the severe sleep deprivation, it was a great experience! I must also thank Santi Serrano for his assistance with PAUS databases and catalogues, as well as Pablo Fosalba. Following with the people at ICE and/or the PAUS collaboration, I must also show my gratitude to Isaac, Arnau and Anna for the encouragement and guidance during the arduous process of searching for a postdoc position, to Gosia for collaborating with the analysis of PAUS SEDs and observing too at the WHT, and also Giorgio for being another great observation partner, as well as a really chill guy.

Regarding other students at ICE, I have to mention David for being a great friend since the master, and for not being the best loser at the game of small rocks, yet playing with me on a regular basis. Also Marina and Andrea for bearing with me during long conversations on the evenings at ICE, and their help with all the paperwork needed to submit this thesis. And of course Benjamin, Cristian and Ivan for all the offline and online discussions about life, science and chinese cartoons, and for providing a resemblance of social life with coworkers in these strange times.

I also need to show my gratitude to all the people from NAOC that helped me and made my year at Beijing an amazing experience. First and foremost to Qi Guo for being my host there, dealing with all the paperwork needed, and also providing recommendation letters.

Moreover, from the CCG department I have to thank as well Shihong and Shuangpeng for their indispensable help finding an apartment and settling there, Huijie for showing me so much stuff and filling proposals with me, Tianchi for inviting me to his wedding, Yingjie for all his assistance with technological stuff, as well as all the conversations during dinner, Kemeng and Qing for the badminton games (although I was quite awful at it), Haonan for teaching me Twilight Struggle (not an easy task!), and especially Jia for all the board games, the evenings playing video games, and the most amazing fight ever witnessed on a Switch. I should also mention some westerners from NAOC here, namely Greg for being a great office mate and for all the coffee breaks, and Enrique and Vero for the nights at that karaoke whose name I cannot even remember, among many other things.

Having mentioned the people I have met during these PhD years (apologies if I forgot anybody!), I should also thank some of the old friends of Castellon, my hometown, for making life a little bit less boring. Without any particular order:

Arturo, Forny, Pablo, Jorge, Abel, Javi, Nahuel, Sergio, Adri, thank you. And also my colleagues from the engineering days: Esteban, Dani, Carlos, Aser, Jordi. Thank you all. I'm glad we still keep in touch after all these years.

Finally, I must conclude by showing my deepest gratitude to my family. None of this would have been possible without them. Thanks to my parents for supporting me and my decisions during all these years, and raise me to be someone able to do this. And thanks to my brother for being there (almost) all this time.

Contents

List of Figures

List of Tables

List of Abbreviations

FLRW	**F**riedmann-**L**emaitre-**R**obertson-**W**alker
ΛCDM	**Λ**(Lambda)-**C**old **D**ark **M**atter
CMB	**C**osmic **M**icrowave **B**ackground
BBN	**B**ig **B**ang **Nucleosynthesis**
2PCF	**2**(Two)-**P**oint Correlation Function
BAO	**B**aryon **A**coustic **O**scillations
Lyα	**L**yman-α(alpha) (emission/absorption line)
Lyβ	**L**yman-β(beta) (emission/absorption line)
IGM	**I**nter**G**alactic **M**edium
SDSS	**S**loan **D**igital **S**ky **S**urvey
FGPA	**F**luctuating **G**unn-**P**eterson **A**pproximation
CCD	**C**harge-**C**oupled **D**evice
SNR	**S**ignal-to-**N**oise **R**atio
IM	**I**ntensity **M**apping
PAUS	**P**hysics of the **A**ccelerating **U**niverse **S**urvey
eBOSS	**e**xtended **B**aryon **O**scillation **S**pectroscopic **S**urvey
DESI	**D**ark **E**nergy **S**pectroscopic **I**nstrument
LRGs	**L**uminous **R**ed **G**alaxie**s**
FWHM	**F**ull **W**idth at **H**alf **M**aximum
NB	**N**arrow **B**and
SED	**S**pectral **E**nergy **D**istribution
AGN	**A**ctive **G**alactic **N**uclei
SFR	**S**tar **F**ormation **R**ate
RA	**R**ight **A**scension
dec	**dec**lination

To my grandmother. Don't worry, I'm doing fine.

Chapter 1

Introduction to standard cosmology

In this introductory chapter we will lay out a very brief history of the development of Cosmology as a fully-fledged physical science, and from there explain in more detail the most basic theoretical foundations of the standard ΛCDM cosmological model: the FLRW metric, the Friedmann equations, and the evolution of energy densities and the scale factor according to them, as well as the different definitions of distance commonly used in cosmology. We will conclude this chapter with a very brief exposition of the evolution of the Universe and its different stages, according to the current paradigm.

1.1 Development of Physical Cosmology

If the broadest definition is to be considered, Cosmology is the study of the origin and evolution of the Universe. When applying the scientific method to this task, some important caveats arise, such as:

- There is only one universe (at least, that we can observe). Therefore, when considering the Universe as a whole, there is only one possible experiment to examine; our Universe as it is. This has crucial implications for both verification of hypothesis via observation and formulation of hypothesis via induction, two keystones of the scientific method. Are these still valid when there is only one observation to be made?

- Other physical sciences deal with the behaviour of systems that are ultimately isolated in both time and space; this means that an arbitrary set of initial conditions may be set for the sake of studying the evolution of the system. However, the initial conditions of the Universe are by definition the origin of everything, and thus when going back in time far enough any assumption strong enough about an initial state enters the domain of metaphysics.

- Since we live in the Universe and it can be observed by us, the nature of the Universe must allow the existence of conscious life. Consequently, it can be argued that a particular feature of the Universe (e.g., energy densities, fundamental constants) is set to a certain value simply because a significant variation in this value would mean that we would not be here to observe it. Are these anthropic reasonings enough of a scientific explanation?

Due to these issues, among others, cosmology remained a purely philosophical discipline during most of Modern history. In fact, this was still a common viewpoint

as late as in the 1960s (Smeenk and Ellis, 2017). However, several advances in the last century, both theoretical and observational, have led to the development of cosmology into a fully recognised physical science. While this science can be called Physical cosmology in order to differentiate it from the broader definition considered before, in this thesis we will always refer to Cosmology exclusively as the scientific study of the Universe.

In this section, we will present a succinct overview of these discoveries that have turned current cosmology into a full branch of physics, with large systematic observation programs, theoretical models capable of elaborating several testable predictions, and measurements of the parameters of these models (sometimes with exquisite precision, e.g., the CMB temperature).

1.1.1 The Cosmological Principle

One assumption that lies at the core of every modern cosmological theory is that our position in the Universe is not special, i,e., not significantly different from any other possible position. While this is a concept that can be originally attributed to Copernicus (the Earth is not the centre of the Universe, but just another planet orbiting the Sun), it can be extrapolated to cosmic scales, with the following implications:

- The Universe is isotropic: its properties are the same regardless of the direction where it is observed.
- The Universe is homogeneous: its properties are the same in all of its points. This is a logical result of total isotropy (since an inhomogeneity would necessarily imply an observable anisotropy for a certain observer).

It is important to note that these two axioms are only applicable to large enough scales; our daily experience shows us inequivocally that, for a scale small enough, the Universe is clearly anisotropic and inhomogenous.

1.1.2 General Relativity

The publication of the theory of general relativity (Einstein, 1915) was a turning point in our understanding of physics, since not only redefines gravitational interactions as a consequence of the geometry of space and time, but also considers time and space as different dimensions of the same four-dimensional manifold (usually called spacetime). As of today, general relativity is the accepted description of gravitation, being an extremely accurate description of gravitational interactions at all macroscopic scales, and standing the test of all the new observations carried out over the last 100 years. In fact, the recent observation of gravitational waves (Abbott et al., 2016) is the most recent verification of a theoretical prediction of general relativity.

Two years after the original publication, Einstein applied the field equations of general relativity to cosmology, with the aim of obtaining a description of the evolution of the Universe as a whole (Einstein, 1917). The resulting equations needed the addition of a constant to keep the universe stationary (since this was the consensus of physicists at the time); the cosmological constant Λ. Some years later, it was observed that the Universe was actually expanding, which made Einstein retract from this cosmological constant. Nevertheless, the cosmological constant Λ is a key feature of the current standard cosmological model, in order to account for the observed expansion.

After the initial work of Einstein, the task of applying general relativity to derive a set of cosmological equations was followed by Alexander Friedmann (Friedmann, 1922; Friedmann, 1924), and independently by Georges Lemaitre (Lemaître, 1931). Afterwards, Howard Robertson and Arthur Walker demonstrated rigorously that the equations derived by Friedmann and Lemaitre were the only possible solution consistent with the cosmological principle, regardless of the field equations of general relativity used to derive them (Robertson, 1935).

Consequently, this solution of the general relativity field equations in accordance with the cosmological principle is known as the Friedmann-Lemaitre-Robertson-Walker metric (FLRW). This FLRW metric will be developed and discussed in section 1.2.1.

1.1.3 Observations of an expanding Universe

One of the facts in cosmology that has been verified the most by different observational probes is that, from our point of view, the Universe is currently ongoing an accelerated expansion. In other words, on large enough scales all objects we can observe are moving away from us, and the farther away they are, the faster they move. This phenomenon was first formally suggested in Lemaître, 1927, and observational confirmation came shortly after.

The first measurement of this accelerated expansion was done by Edwin Hubble (Hubble, 1929), who used to the relationship between period and luminosity in variable Cepheid stars to measure the distance to close galaxies. By observing the period of the Cepheid stars contained in these galaxies, its actual luminosity can be inferred, and by comparing the actual luminosity to the observed luminosity, the distance to these Cepheid stars (and thus, the galaxies containing them) can be computed.

Hubble combined these distance measurements with the redshift of these same galaxies observed by Vesto Slipher (Slipher, 1913), which were a measure of the relative velocity of the galaxies with respect to us. This study showed that there was a strong correlation between distance and velocity (Figure 1.1).

The slope in the linear fit shown in fig. 1.1 is the value of the Hubble constant. For a more thorough explanation of this constant, as well as the relationship between redshift and velocity/distance, see section 1.2.6. This experimental confirmation of the accelerated expansion challenged the conception of a stationary Universe that was widely held at the time, diplaying it as a dynamic system that was currently evolving.

It is important to note how revolutionary this notion was back then; not even a decade before, it was discussed if the galaxies observed by Hubble (called *nebulae* at that point) were small stellar aggregations or gas clouds embedded into the Milky Way or galaxies like our own, much farther away than any object ever observed before. This debate was actually held in 1920 between two astronomers, Harlow Shapley (defending that the nebulae were part of our galaxy) and Heber Curtis (proposing that they were independent galaxies).

Ever since the pioneering observations by Hubble, the accelerated expansion of the Universe and the value of the Hubble constant have been determined by a myriad of other cosmological probes. One of the most important probes will be briefly discussed in the next subsection.

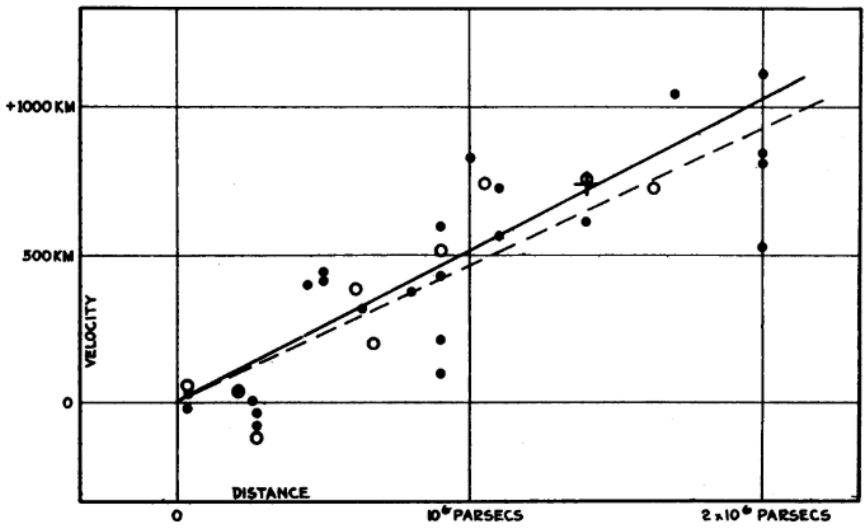

FIGURE 1.1: Original plot in Hubble, 1929, displaying his distance measurements in the x axis vs the velocities inferred from Slipher's redshift measurements. Note that the units in the y axis should be velocity (km/s), not distance (km). Also, an error in distance calibration yields a slope almost an order of magnitude larger than current accepted values.

1.1.4 The Cosmic Microwave Background and the Big Bang theory

With solid evidence that the Universe was currently expanding, and theoretical models that supported this expansion, several questions aroused regarding the origin and evolution of the Universe. Has expansion being going on forever? If so, how was the Universe at earlier stages?

In order to answer these questions, two models were proposed: the steady-state and the Big Bang. The steady-state model proposed that the Universe was forever expanding yet effectively stationary; a continuous creation of matter was introduced in Einstein's equations, which cancelled out the decrease of density due to expansion (Hoyle, 1948). This made the expansion compatible with the notion of an eternal and immutable Universe.

On the other hand, the Big Bang model proposed that the Universe expanded forever, but this implied that at as we went back time, the Universe necessarily was denser and smaller (and thus hotter, since the energy density would be increased). By reasoning in this manner, it followed that long ago enough the Universe was a single point of infinite density; a singularity where the concepts of space and time lost their meaning (). Unlike the state-steady model, the Big Bang model implied that the Universe was fundamentally different before, and that its age was finite, since the initial singularity was its inception. (Ironically, the term "Big Bang" was coined by Fred Hoyle, one of the main proponents of the steady-state model).

A fortunate detection in 1964 provided decisive evidence for the Big Bang model, which has been the consensus among cosmologists ever since. While working with a highly sensitive radio antenna designed to detect satellite signals, American physicists Arno Penzias and Robert Wilson discovered a uniform signal that could not be accounted for, and remained the same regardless of the time or the day or the region of the sky that was being observed. Given its immutability, such a radiation necessarily had its origin in extragalactic sources.

This was the Cosmic Microwave Background (CMB); the first electromagnetic radiation detectable after the Big Bang. Since the Universe expands, objects farther away are more redshifted (and older, since the radiation we observe has taken a longer time to arrive). Assuming then that the Universe is of finite age, there should be an extremely redshifted first radiation, well below the optical spectrum. This is exactly what Penzias and Wilson first observed, and although their initial temperature estimate was higher, nowadays it is known to correspond to a thermal blackbody emission of $T \sim 2.725K$.

Strictly speaking, the CMB is not a signal of the very beginning of the Universe; it is the first radiation at the epoch of recombination, when the Universe became cold enough to allow electrons and protons to combine into hydrogen atoms. Prior to this point, both baryonic matter and photons where coupled in a plasma state, where the photons where continuously scattered. When recombination happened, photons became able to freely travel without scattering, which made the Universe effectively transparent to the CMB radiation we currently see.

Although the CMB seemed completely homogeneous at first, theory predicted small inhomogeneites of the order of $10^{-4} - 10^{-5}$ (e.g., Peebles and Yu, 1970). This apparent homogeneity is due to the fact that, just an instant before, the Universe was a dense plasma, and no time had passed to allow for the formation of structure we see today. The anisotropies of the CMB were first studied in detail by the COBE mission, which measured the CMB variations in the full sky (Smoot et al., 1992), and have turned out to be one of the most valuable probes to constrain cosmological models. Later space missions have again observed the CMB with higher resolution, namely the Wilkinson Microwave Anisotropy Probe (WMAP, Spergel et al., 2003) and Planck (Planck Collaboration et al., 2016). Figure 1.2 shows a combination of the CMB maps obtained by these three probes.

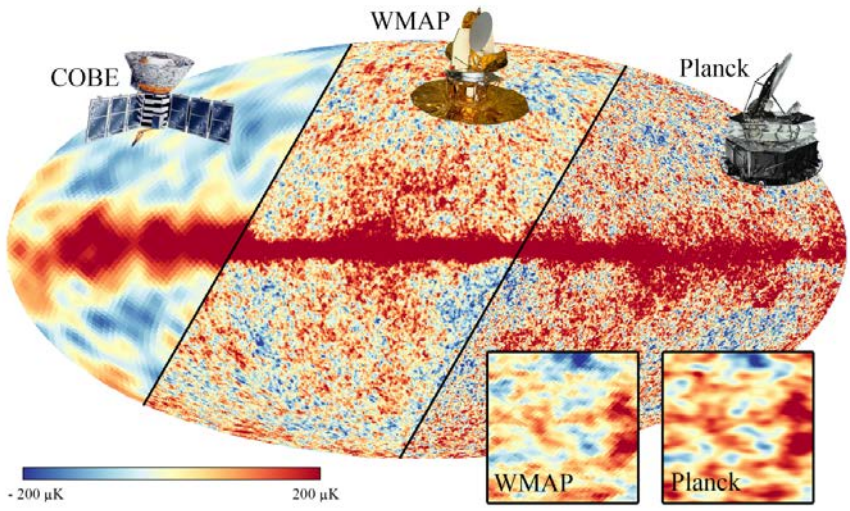

FIGURE 1.2: Combined maps of CMB by the space missions COBE, WMAP and Planck. Angular resolution is noticeably increased in each survey. Credit: J. Gudmunsson.

1.2 The ΛCDM Model

The current standard model in cosmology is commonly known as the ΛCDM model, where Λ stands for the Dark Energy in the form of a cosmological constant (originally named Λ by Einstein), and CDM is an acronym for Cold Dark Matter. While this model fails to account for some observed features in the universe, such as the disparity in the Hubble constant measurements (Bernal, Verde, and Riess, 2016) or the anomalies in the CMB (Schwarz et al., 2016), it is still considered a massive success, which accounts for most of the cosmological observations carried out in the last century with relative simplicity.

In this subsection we will show a small, non-extensive overview of the aspects of the model more related to the work of the thesis. This does not intend to cover the entirety of the model, nor even to describe all of the parameters that compose it; for a complete and comprehensible rundown of the standard model of cosmology, we refer the reader to Dodelson and Schmidt, 2020, chapters 1 and 2.

1.2.1 The scale factor and the FLRW metric

By assuming the cosmological principle (homogeneity and isotropy at large enough scales), we already place very stringent constraints on the geometry of the Universe. One of the consequences of this principle is that, if space is expanding, as multiple observational evidence shows, it must do so at the same rate in all points. Therefore, the expansion (or contraction) of space can be simply modelled with a single parameter dependant on time: the scale factor $a(t)$.

Assuming the simplest case of an Euclidean three-dimensional space, we can define two points with Cartesian coordinates $\boldsymbol{x}$ and $\boldsymbol{x} + d\boldsymbol{x}$. For a given time t, the distance vector $d\boldsymbol{r}$ between these two points will be:

$$d\boldsymbol{r} = a(t)d\boldsymbol{x}. \tag{1.1}$$

While the actual, physical distance $\boldsymbol{r}$ will evolve with time according to $a(t)$, it is important to note that the vector defined by the difference between the Cartesian coordinates, $d\boldsymbol{x}$ will remain constant. Since the scale factor is simply a re-scaling, it will be dimensionless, and thus $d\boldsymbol{x}$ will also have length units. Therefore, it can also be considered a distance, although not physical.

This invariant distance $d\boldsymbol{x}$ is known as the *comoving distance*, and analogously the coordinates $\boldsymbol{x}$ are called *comoving coordinates*. These two concepts are extensively used in cosmology, given that the independence from the scale factor $a(t)$ allows to easily perform calculations and use expressions valid at any point of cosmic time.

By convention, the scale factor is set to 1 at the present time t_0, so in our local environment comoving and physical distances are the same ($a(t_0) = 1$). Given that the Universe has been expanding ever since the beginning, $a(t < t_0) < 1$ at any point in the past.

The application of the scale factor can be formally described with a metric, according to the framework of general relativity: the FLRW metric. First, let us start with the general definition of a metric:

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu, \tag{1.2}$$

where ds^2 is the square of the length element (also known as *line element*), dx^μ and dx^ν are the elements μ and ν of the difference between the coordinates of two points

$\boldsymbol{x}$ and $\boldsymbol{x} + d\boldsymbol{x}$, and $g_{\mu\nu}$ is the metric itself. Therefore, the metric is the relationship between difference in coordinates and distance.

Since we are considering general relativity, the spacetime where this metric (and the coordinates) are defined is a 4-dimensional manifold, where by convention the first coordinate ($\mu = 0$) is time, and the other three are the spatial coordinates. For the simplest possible case, a flat spacetime devoid of matter, the metric is a tensor with the following components

$$g_{\mu\nu} = \begin{pmatrix} -c^2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{1.3}$$

Where the first element of the diagonal is g_{00}, relative exclusively to the time coordinate, and the other elements of the diagonal are the spatial elements of the metric, g_{11}, g_{22}, g_{33}. The absolute value of g_{00} is c, the speed of light. Following eq. (1.1), the scale factor $a(t)$ can be implemented into the metric simply by replacing the spatial component, which yields

$$g_{\mu\nu} = \begin{pmatrix} c^2 & 0 & 0 & 0 \\ 0 & -a(t)^2 & 0 & 0 \\ 0 & 0 & -a(t)^2 & 0 \\ 0 & 0 & 0 & -a(t)^2 \end{pmatrix}. \tag{1.4}$$

This is the tensor of the FLRW metric in a flat spacetime. Applying it to eq. (1.2) and expanding, we obtain

$$ds^2 = c^2 dt^2 - a(t)^2 d\boldsymbol{x}, \tag{1.5}$$

where $d\boldsymbol{x}$ are the spatial components of the coordinate tensor, and dt is its time component. This is the FLRW metric for a flat Universe. Given that isotropy is always assumed for this metric, it is often more convenient to express the metric in spherical coordinates as

$$ds^2 = c^2 dt^2 - a(t)^2 \left(d\chi + \chi^2 d\Omega^2 \right), \tag{1.6}$$

where χ is the radial component of the spatial coordinates, and $d\Omega$ the solid angle component, which comprises both the polar component θ and azimuthal component ϕ (usually, by convention $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$).

If the Universe has an intrinsic spatial curvature (which does not violate the cosmological principle as long as this curvature is uniform), the FLRW metric adopts a more complex form, and eq. (1.6) is then expressed as:

$$\begin{aligned} ds^2 &= c^2 dt^2 - a(t)^2 \left(d\chi + k^{-1} \sin^2(\chi\sqrt{k}) d\Omega^2 \right) \quad \text{if} \quad k > 0 \\ ds^2 &= c^2 dt^2 - a(t)^2 \left(d\chi + |k|^{-1} \sinh^2(\chi\sqrt{|k|}) d\Omega^2 \right) \quad \text{if} \quad k < 0. \end{aligned} \tag{1.7}$$

In this case, k is the curvature, with units of length^{-2}. If the curvature is larger than 0, the universe has spherical geometry (also known as *closed universe*), and if the curvature is smaller than 0, the geometry is hyperbolic (*open universe*). We will not develop more these equations for a Universe with curvature, since observational evidence points to a flat or almost-flat Universe (e.g., Planck Collaboration, 2020).

For this very same reason, a flat Universe will be assumed for the entirety of this thesis.

1.2.2 The Friedmann equations

So far in this section, we have briefly explained the FLRW metric, which is the most general set of geometrical constraints that satisfy the cosmological principle. In this metric, the only time-dependant parameter is the scale factor $a(t)$; if a flat Universe is assumed, it is also the only free parameter of the metric.

Therefore, the capability to explain and predict the evolution of $a(t)$ is an essential requirement for any cosmological model. Following general relativity, we can find a relationship between $a(t)$ and the content of the Universe; the Friedmann equations. For convenience, we will assume in this subsection that $c = 1$ (natural units)

First, let us consider the Einstein equation, which links the matter and energy content of spacetime with its geometry:

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} - \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{1.8}$$

Here $g_{\mu\nu}$ is the metric tensor as in eq. (1.2), G is the Newtonian constant of gravitation, Λ the cosmological constant alreadyu discussed in section 1.1.2, $T_{\mu\nu}$ the energy-momentum tensor, and $R_{\mu\nu}$ is the Ricci tensor (R is its contraction, also known as Ricci scalar: $R = g^{\mu\nu}R_{\mu\nu}$). The Ricci tensor is a mathematical object dependant solely on the metric and its derivatives:

$$R_{\mu\nu} = \frac{\partial \Gamma^{\lambda}_{\lambda\mu}}{\partial x^{\nu}} - \frac{\partial \Gamma^{\lambda}_{\mu\nu}}{\partial x^{\lambda}} + \Gamma^{\lambda}_{\mu\sigma}\Gamma^{\sigma}_{\nu\lambda} - \Gamma^{\lambda}_{\mu\nu}\Gamma^{\sigma}_{\nu\sigma}, \tag{1.9}$$

where $\Gamma^{\lambda}_{\mu\nu}$ is the Christoffel symbol of second kind:

$$\Gamma^{\lambda}_{\mu\nu} = \frac{1}{2}g^{\lambda\kappa}\left(\frac{\partial g_{\kappa\mu}}{\partial x^{\nu}} + \frac{\partial g_{\kappa\nu}}{\partial x^{\mu}} - \frac{\partial g_{\mu\nu}}{\partial x^{\kappa}}\right). \tag{1.10}$$

Regarding $T_{\mu\nu}$, it is a symmetric tensor whose components represent the following (with units of energy density):

- T_{00}: Energy density
- T_{0i},T_{i0}: Energy flux/momentum density
- T_{ii}: Pressure
- T_{ij},T_{ji} with $i \neq j$: Momentum flux/shear stress

Here we have adopted the convention of using Latin letters when referring strictly to the spatial components ($x > 0$) and Greek letters when referring to any component including time ($\mu \geq 0$).

The Einstein equation is indeed a extremely complex problem on itself, but fortunately the FLRW metric simplifies it to the point where an exact solution can be found. If we compute the Ricci tensor using this metric, the Christoffel symbols with two or three temporal indices ($\mu = 0$) vanish, which yields only two kinds of non-null components of the Ricci tensor: the strictly temporal component R_{00} and the spatial components R_{xy}. For the general case of FLRW metric with curvature, these non-null components are:

$$R_{00} = -3\frac{\ddot{a}}{a}, \tag{1.11}$$

$$R_{ij} = -\frac{g_{ij}}{a^2}\left(2k + 2\dot{a}^2 + a\ddot{a}\right), \tag{1.12}$$

where the dots in $\dot{a}$ and $\ddot{a}$ represent the first and second time derivatives respectively, and k is the curvature of the metric. The expression of the Ricci scalar for these components of the tensor and the FLRW metric is:

$$R = -6\frac{\ddot{a}}{a} - 6\left(\frac{\dot{a}}{a}\right)^2 - 6\frac{k}{a^2} \tag{1.13}$$

Now, in order to satisfy the cosmological principle, we assume the most general form of an energy-momentum tensor corresponding to a perfect, isotropic fluid:

$$T_{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & -p & 0 & 0 \\ 0 & 0 & -p & 0 \\ 0 & 0 & 0 & -p \end{pmatrix}. \tag{1.14}$$

Where ρ is its density, and p its pressure. Note that the non-diagonal components of the tensor can be interpreted as either energy fluxes or momentum fluxes; a net flux in any direction would result in an anisotropy, and thus a violation of the cosmological principle. Under these assumptions, the temporal component of the Einstein equation is

$$3\left(\frac{\dot{a}}{a}\right)^2 + 3\frac{k}{a^2} - \Lambda c^2 = 8\pi G\rho, \tag{1.15}$$

which can be expressed as

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda}{3} - \frac{k}{a^2}. \tag{1.16}$$

This is a differential equation that constrains the evolution of $a(t)$; the first Friedmann equation, since it establishes a relationship between the velocity of the scale factor evolution and the physical properties of the Universe.

The second Friedmann equation is derived from the spatial part of eq. (1.8). Replacing eq. (1.12) and eq. (1.13) in eq. (1.8), we find for the spatial components of the trace ($i = j$):

$$2\frac{\ddot{a}}{a} + \left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} - \Lambda = -8\pi G p \tag{1.17}$$

Where we have already divided by the metric, g_{ii}. We divide by 2 and rearrange the equation in the following form

$$\frac{\ddot{a}}{a} + \frac{1}{2}\left(\frac{\dot{a}}{a}\right)^2 = -4\pi G p + \frac{\Lambda}{2} - \frac{k}{2a^2}. \tag{1.18}$$

If we subtract to this equation the first Friedmann equation (eq. (1.16)) divided by two, the expression we obtain is the second Friedmann equation:

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) + \frac{\Lambda}{3}. \tag{1.19}$$

This expression is also a differential equation, independent of the first. In this case, this equation sets a relationship between the acceleration of the scale factor and the physical properties of the Universe. In the following subsection, we will see how these equations are commonly expressed in terms of observational parameters with a more specific interpretation.

1.2.3 The Hubble constant and the deceleration parameter

As explained in section 1.1.3, the Hubble constant is the relationship between the distance of an object and the velocity it drifts away from us. It can simply be defined as:

$$v = H_0 d. \tag{1.20}$$

Where v is the velocity of the object (drifting away from us), H_0 is the Hubble constant and d is its physical distance. The velocity v in this expression is due to the expansion of the Universe (also known as Hubble flow); however, the actual observed velocity of an object will not only have this Hubble flow component. Other velocities (*proper motions*), due to the relative movement of galaxies respect to other galaxies or larger structures such as clusters, will also contribute to the net movement of the observed object, biasing the measurement of H_0. Therefore, systematic observations of large cosmic volumes are needed to accurately determine its value.

Figure 1.1 shows the original measurement of H_0, but not only it is an observation of an extremely limited sample; distances are not properly measured due to a error in the calibration of variable Cepheid stars, and thus the resulting H_0 is almost an order of magnitude larger than current observed values. In comparison to this first plot, fig. 1.3 shows a compilation of measurements of H_0 from different observational probes over the last 20 years.

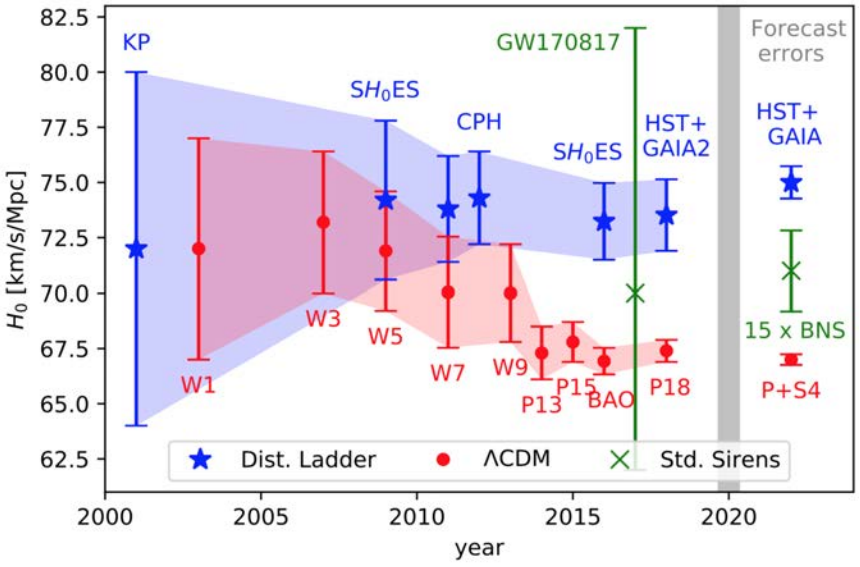

FIGURE 1.3: Values of H_0 determined by different observational probes, ordered chronologically by date of publication. Blue values represent local-Universe measurements based on calibration of Cepheid stars, red values are based on CMB observations, and green values are direct measurements using standard sirens. Obtained from Ezquiaga and Zumalacárregui, 2018, Figure 8.

By looking at fig. 1.3, it is evident that this last decade a clear tension has appeared between observational probes. Local measurements of H_0 yield higher values than observations based on the early-Universe data from CMB, and the increase in precision with better observations has worsened this tension. Since this uncertainty in H_0 is crucial when defining cosmological quantities, it is common practice to define H_0 as a function of the h parameter:

$$H_0 \equiv 100\,h \quad (\mathrm{Km\,s^{-1}\,Mpc^{-1}}) \tag{1.21}$$

This h parameter is generally accepted to be in a range $0.6 < h < 0.8$, and expressing cosmological distances or volumes in units dependant on h is a convenient way of computing and expressing data while always considering this uncertainty in an implicit manner.

The scientific value of H_0 lies not only in being the first observation of an expanding Universe, but in its direct relation with the first Friedmann equation, which puts constraints on the physical properties of the Universe. In order to show this, first let us assume a value of the Hubble constant as a function, of cosmic time, $H(t)$. This time-dependent Hubble constant is also known as the *Hubble parameter*; in fact the Hubble constant is simply the Hubble parameter at present time ($H_0 = H(t_0)$). it is important to note that, despite being dependant on time, in order to satisfy the cosmological principle $H(t)$ must be equal at all points in space.

Now, let us consider an observer at time t, and an object separated from the observer a physical distance dr (since we assume the Universe isotropic, the distance can simply be expressed as an scalar). If we note its comoving distance as $d\chi$, and assume that the observed velocity of the object corresponds to the Hubble flow (negligible proper motion), the physical distance and observed velocity v can be written as

$$dr = a(t)d\chi \quad ; \quad v = \frac{dr}{dt} = \frac{da(t)}{d(t)}d\chi + a(t)\frac{d\chi}{dt} = \dot{a}(t)d\chi. \tag{1.22}$$

Since proper motion is assumed negligible, all the observed velocity must come from the Hubble flow (the variation of the scale factor), and thus $d\dot{\chi} = 0$. With these expressions, eq. (1.20) can be redefined as

$$\dot{a}(t)d\chi = H(t)a(t)d\chi, \tag{1.23}$$

and thus the expression of the Hubble parameter as a function of the scale factor is

$$H(t) = \frac{\dot{a}(t)}{a(t)} \tag{1.24}$$

This expression of the Hubble parameter is the square root of the left-hand side of the first Friedmann equation (eq. (1.16)), which can be written as

$$H^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda}{3} - \frac{k}{a^2}, \tag{1.25}$$

Which links directly the value of the Hubble constant at a given cosmic time with the energy density of the Universe, as well as the possible values of its curvature and cosmological constant.The second Friedmann equation is also usually expressed as a function of another variable, the deceleration parameter

$$q(t) \equiv \frac{-\ddot{a}(t)a(t)}{\dot{a}(t)^2}, \tag{1.26}$$

which represents minus the acceleration of the scale factor, in dimensionless units (and its value at current time is noted q_0). Using this deceleration parameter as well as eq. (1.24), eq. (1.19) can be expressed as

$$q = \frac{4\pi G}{3H}(\rho + 3p) - \frac{\Lambda}{3H}. \tag{1.27}$$

1.2.4 Energy content of the Universe

Once we have managed to define a clear relationship between an observable such as the Hubble parameter, and the density of energy content in the universe, the question that remains is, what is the physical meaning of this density ρ, and how does it evolve with a? In this subsection we will deal with these questions.

Matter density

According to the mass-energy equivalence from general relativity, the energy contained of a certain amount of matter of mass m is

$$E = mc^2, \tag{1.28}$$

which makes its energy directly proportional to its mass. Assuming matter conservation on cosmological scales, the matter energy density in a certain point will solely depend on the volume; if the universe expands, the matter energy density will decrease, since the same energy will be contained in a larger volume. Given that space is three-dimensional, an arbitrary physical volume is going to be proportional to the cube of the scale factor, so a matter energy density will be proportional to its inverse:

$$\rho_{\mathrm{m}}(t) \propto a(t)^{-3}. \tag{1.29}$$

Knowing that by convention $a(t_0) = 1$, and noting the current matter density as $\rho_{\mathrm{m}0}$, the relation between the current matter density and the matter density at any previous time is

$$\rho_{\mathrm{m}} = \rho_{\mathrm{m}0} a^{-3} \tag{1.30}$$

This density includes both the matter we can see and interact with (baryonic matter), as well as the dark matter, which is hypothesised to exist due to several independent observational probes (e.g., galaxy velocity curves, weak lensing), but whose nature is unknown.

There are essential differences between baryonic matter and most models of dark matter; namely, dark matter is assumed to interact only through the gravitational force, thus being collisionless and unable to emit or react to electromagnetic radiation. However, on the scales where the cosmological principle holds (and therefore, all the equations of this section), gravity is the only non-negligible force, and thus both kinds of matter can be considered the same.

Radiation density

Another component of the total energy budget of the Universe is radiation. However, photons are massless, and its energy depends on its wavelength:

$$E = \frac{h_P c}{\lambda}. \tag{1.31}$$

Here, h_P is the Planck constant (noted with a subscript to differentiate it from the h parameter), and λ the wavelength of the photon. This has a key implication: when the Universe expands, not only the photon density will be reduced because of the increase in volume, but the energy of each photon will also decrease, given that their wavelength will expand proportionally to a.

Since the photon density is proportional to the inverse of the cube of a, and their energy is proportional to the inverse of a, their energy density will be proportional to the negative fourth power. Therefore,

$$\rho_{\mathrm{r}} = \rho_{\mathrm{r}0} a^{-4} \tag{1.32}$$

Dark energy and the equation of state

The cosmological constant, as displayed in eq. (1.16), can be thought as an energy density completely independent of $a(t)$; in other words, an energy density absolutely constant in time. The energy density due to the cosmological constant ρ_Λ needs have the same units as ρ to be added as a part of the total energy density budget, which can be done with the following expression

$$\rho_\Lambda = \frac{\Lambda}{8\pi G}, \tag{1.33}$$

and considering that Λ is constant, its dependency with time can simply be written as

$$\rho_\Lambda = \rho_{\Lambda 0} \tag{1.34}$$

The physical interpretation of this last equation is that Λ represents a energy density intrinsic of space, the *vacuum energy* (since even if space is completely empty, this energy content is still there). In cosmology, this phenomenon is called dark energy. It is needed to explain the observed accelerated expansion of the Universe with Friedmann equations, but its physical nature remains unknown.

While the cosmological constant/vacuum energy is the standard approach to model dark energy (hence the Λ from ΛCDM), other possible forms of dark energy are also considered. The behaviour of this different types of dark energy is usually described by the equation of state, usually expressed as:

$$p = \omega \rho. \tag{1.35}$$

This equation expresses the pressure of a fluid as a function of its density. The evolution of the energy density as a function of the scale factor can be derived from the continuity equation

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + p) = 0, \tag{1.36}$$

which states that the variation of energy density in a point must be equal to the flux of energy density in that point (under the assumption that on cosmological scales all fluxes are due to the variation of a). If we substitute eq. (1.35) in this differential equation, we find the following solution for an arbitrary fluid:

$$\rho = \rho_0 a^{-3(1+\omega)} \tag{1.37}$$

Depending on the value of ω, different models of dark energy are defined:

- $\omega = -1$: Vacuum energy/cosmological constant.
- $\omega = \omega(a)$: Variable dark energy. Most models correspond to a scalar field (*quintessence*). Current values need to be below -1/3 to be in accordance with the observed accelerated expansion.
- $\omega < -1$: Phantom energy. Although the Friedmann equations can yield solutions with these models, they imply that vacuum energy is unstable, so they are less considered.

The ω parameter and the equation of state can also be used to model the behaviour of not only dark energy, but also the other components of the Universe. For matter, $\omega = 0$ (pressureless fluid), and for radiation $\omega = 1/3$ (radiation pressure). Replacing these values of ω in eq. (1.37), we recover the expressions in eq. (1.30) and eq. (1.32).

1.2.5 Evolution of energy densities and the Hubble parameter

Having described the evolution of the energy densities for each component of the Universe as a function of a, eq. (1.25) can be rewritten as:

$$H^2 = \frac{8\pi G}{3}\left(\rho_{\mathrm{m}0}a^{-3} + \rho_{\mathrm{r}0}a^{-4} + \rho_{\Lambda 0}\right) - \frac{\mathrm{k}}{a^2}. \tag{1.38}$$

Following the same reasoning as dark energy, the curvature term in the right-hand side of the equation can be expressed as an energy density, $\rho_{\mathrm{k}0} = k$, such that

$$H^2 = \frac{8\pi G}{3}\left(\rho_{\mathrm{m}0}a^{-3} + \rho_{\mathrm{r}0}a^{-4} + \rho_{\Lambda 0} + \rho_{\mathrm{k}0}a^{-2}\right). \tag{1.39}$$

In this case, $\rho_{\mathrm{k}0}$ does not have a physical meaning such as vacuum energy for ρ_Λ, since k is an strictly geometrical term. This is simply a convention to write this expression of the Hubble constant more elegantly, and express the effect of the curvature on the evolution of the scale factor in the same terms as any other component of the Universe.

All these energy densities can be expressed as a function of the current total energy density of the Universe: the critical density ρ_{c}. By convention, $a(t_0) = 1$, and thus if we consider in eq. (1.39) the sum of all energy density values in t_0 equal to the critical density, we can define this critical density like

$$\rho_{\mathrm{c}} \equiv \frac{3H_0^2}{8\pi G}. \tag{1.40}$$

For each one of the components, the dimensionless energy density parameter Ω is defined as the ratio of their current densities and the critical density:

$$\Omega_{\mathrm{m}} \equiv \frac{\rho_{\mathrm{m}0}}{\rho_{\mathrm{c}}}; \quad \Omega_r \equiv \frac{\rho_{\mathrm{r}0}}{\rho_{\mathrm{c}}}; \quad \Omega_\Lambda \equiv \frac{\rho_{\Lambda 0}}{\rho_{\mathrm{c}}}; \quad \Omega_{\mathrm{k}} \equiv \frac{\rho_{\mathrm{k}0}}{\rho_{\mathrm{c}}} \tag{1.41}$$

The fraction of the energy density of the Universe for each component is represented in fig. 1.4, in two plots for different timescales: before CMB and after CMB. The framework of this thesis only is focused in the latter, so neutrinos have not been included in section 1.2.4, since its contribution is only significant at very early times. It is important to note how the contribution of radiation was only non-negligible well before 1 Gyr, and during most of its lifespan the Universe has been dominated

TABLE 1.1: ΛCDM parameters used to plot fig. 1.4, extracted from Planck Collaboration et al., 2016. The parameters σ_8 and n_s will be discussed in section 2.2.2.

Parameter	Value
Ω_Λ (dark energy density parameter)	0.6911 ± 0.0062
Ω_m (matter density parameter)	0.3089 ± 0.0062
$\Omega_b h^2$ (baryonic matter density parameter)	0.02230 ± 0.00014
$\Omega_{cdm} h^2$ (cold dark matter density parameter)	0.1188 ± 0.0010
H_0 (Hubble constant)	67.74 ± 0.46
σ_8 (standard deviation of density fluctuations at 8 Mpc/h, $z = 0$)	0.8159 ± 0.0086
n_s (primordial power spectrum spectral index)	0.9667 ± 0.0040

by matter. However, dark energy has slowly become more and more significant, until being the dominant component in the last ∼4 Gyr. The parameters used to plot fig. 1.4 are displayed in table 1.1.

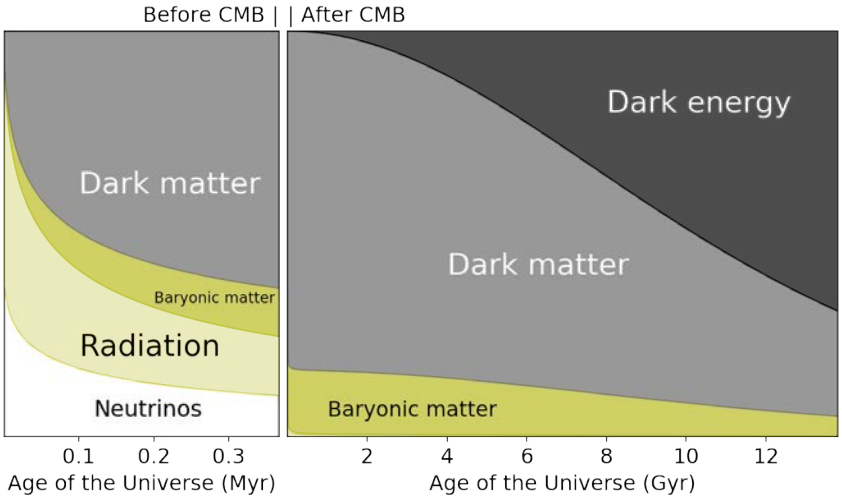

FIGURE 1.4: Fractions of each component of the energy density of the Universe versus time. The discontinuity before and after CMB is due to the change of time scales; the time units in the plot after CMB are three orders of magnitude larger than the plot before CMB, and thus the transition to negligible radiation and neutrino energy densities is too fast to be appreciated in Gyr units.

Combining eq. (1.40) and eq. (1.41) with eq. (1.39), the Hubble parameter can be expressed as a function of the scale factor:

$$H(a) = H_0 \sqrt{\Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda}. \tag{1.42}$$

This expression is used regularly for all cosmological computations of time and distances, as it will be seen in section 1.2.7.

1.2.6 Redshift

In the broadest sense of the term, *redshift* is the difference between the wavelength of the radiation emitted by a source, and the wavelength of this same radiation observed by a certain observer. Its mathematical definition is

$$z \equiv \frac{\lambda_o - \lambda_e}{\lambda_e}, \tag{1.43}$$

where λ_e is the wavelength at the source of emission, and λ_o is the observed wavelength. Usually, this phenomenon is called redshift when $\lambda_o > \lambda_e$ (the observed wavelength is larger than the emission, and thus the observed spectrum has shifted to the red). If $\lambda_o < \lambda_e$, following the same logic is is usually referred to as *blueshift*, although this is never the case for cosmological distance large enough (as in this work).

There may be different physical phenomena producing redshift, such as the relative velocity between the observer and the source (Doppler effect), the difference in gravitational potential between source and observer (gravitational redshift), or the expansion of the Universe (cosmological redshift).

In this subsection we will only discuss cosmological redshift and the Doppler effect. We also note that while other physical mechanisms due to the medium where radiation is transmitted may result in a shifting to the red of the spectrum (e.g., scattering caused by dust), these are usually referred to as *reddening*.

Doppler effect

Let us assume the source and the observer moving away from each other with a relative velocity v; this velocity is simply assumed negative if they are approaching. Due to this relative movement, the distance between two consecutive observed peaks separated in time Δt (the period of the wave) will be increased by $v\Delta t$.

If this hypothetical wave corresponds to electromagnetic radiation, it will always move at speed c, and its period will be $\Delta t = \lambda_e/c$ (since the period is the time required to for the wave to advance a full wavelength). Consequently, the observed and emitted wavelength are related by

$$\lambda_o = \lambda_e + v\lambda_e/c, \tag{1.44}$$

and applying this equality to eq. (1.43), we find

$$z = \frac{v}{c}. \tag{1.45}$$

This equation corresponds to the classical Doppler effect. It is worth nothing that this is a linear approximation that holds for speeds not too close to c. For the cosmological case, where the relative velocities are caused mostly by the Hubble flow, eq. (1.45) stops being valid roughly at $z \sim 0.5$.

Cosmological redshift

The expansion of the Universe has a similar effect on the observed wavelength, if the observed radiation has travelled during enough time so the scale factor a has changed significantly. Let us assume an electromagnetic wave, emitted by a distant object at time t, and observed currently, at t_0. The scale factor factor at the time of emission will be $a = a(t)$, while at the time of observation $a(t_0) = 1$.

Given that the wavelength of the emitted radiation is a physical distance, it will expand according to the increase of the scale factor with time. Therefore, the relation between the emitted and the observed wavelength will be

$$\lambda_e = a\lambda_o. \tag{1.46}$$

Replacing this in eq. (1.43), we find the following equations linking scale factor and redshift:

$$z = \frac{1-a}{a}; \qquad a = \frac{1}{1-z}. \tag{1.47}$$

These straightforward relationships make cosmological redshift more used than the actual scale factor both in equations and data plots. For example, eq. (1.42) is more commonly expressed as

$$H(z) = H_0\sqrt{\Omega_r(1+z)^4 + \Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}. \tag{1.48}$$

Like the scale factor, a certain value of cosmological redshift unequivocally points to a certain time in cosmic history (as well as a distance, as will be explained in the next subsection), and unlike the scale factor, redshift can be directly measured. If the spectrum of an object is observed, by assuming some spectral features of this object known (e.g., emission lines, spectral breaks), the redshift of the object can be determined. The most common observational approaches (photometric and spectroscopic) to determine the redshift of observed objects with optical telescopes will be briefly discussed in chapter 3.

However, these direct observations also include the redshift due to the Doppler effect (peculiar velocity, i.e., the proper motion of galaxies regardless of Hubble flow). While its effect is usually smaller, it is still a source of error that needs to be accounted for. Given that the peculiar velocities are usually far below c, the classical Doppler approximation (eq. (1.45)) holds true for them.

Regarding the recession velocity derived from cosmological redshift, the simple classical eq. (1.45) is only valid at small redshifts, given that redshift is caused by the expansion of the Universe at every point of the trajectory of the redshifted photon, not only the relative motion due to the Hubble flow. The proper expression to compute the cosmological recession velocity is (Davis and Lineweaver, 2001)

$$v(z) = \dot{a}(z)\int_0^z \frac{cdz'}{H(z')}, \tag{1.49}$$

where z' is the integration variable over the specified redshift ranges. This expression yields $v > c$ for $z \gtrapprox 1.5$ in any cosmology consistent with current observations, which at first sight might seem impossible. According to general relativity, c is the asymptotic speed limit of any massive object, regardless of the reference frame where it is observed. However, superluminal velocities do not violate this principle; the reason for this will be briefly explored in section 1.2.7.

1.2.7 Cosmological distances

So far, we have mentioned two definitions of distance in cosmology: the physical distance between objects (also known as *proper distance*), and the comoving distance, which is independent of the scale factor. However, other types of distance are also defined and commonly used for different purposes.

Comoving distance

Working with physical distances in a Universe where the scale factor evolves with time is unnecessarily complex; for most purposes in cosmology, whenever a set of spatial coordinates needs to be used, these are in comoving space.

Given that the scale factor does not modify measured angles (since angles are dimensionless, and the scale factor affects equally all dimensions of space), if spherical coordinates are used the only comoving distance needed to locate an object in comoving space is its radial distance.

Let us consider an observer at the origin of our comoving coordinate system, and at current time t_0, and a distant object that has emitted light at a time t_e. In a infinitesimal time interval dt, light will have travelled a physical distance

$$cdt = a(t)d\chi, \tag{1.50}$$

where $d\chi$ is the comoving distance. This can be easily derived from eq. (1.6) assuming the distance completely radial, and knowing that in general relativity the spacetime interval between two points travelled at light speed is null.

This expression can be integrated from the time of emission to the time of observation, in order to determine the total comoving distance travelled by light, and thus the distance to the observed object:

$$\chi(t_e) = \int_{t_0}^{t_e} dt \frac{c}{a(t)}. \tag{1.51}$$

Applying the change of variable $\dot{a} = da/dt$, and the definition of the Hubble parameter (eq. (1.24)), χ can be expressed as a function of the scale factor:

$$\chi(a) = \int_1^{a_e} da \frac{c}{a\dot{a}} = \int_1^{a_e} da \frac{c}{a^2 H(a)}. \tag{1.52}$$

Finally, using the relationship between scale factor and redshift (eq. (1.47)), and the expression of the Hubble parameter as a function of redshift (eq. (1.48)), we find

$$\chi(z) = \frac{c}{H_0} \int_0^{z_e} \frac{dz}{\sqrt{\Omega_r + \Omega_m(1+z)^{-1} + \Omega_k(1+z)^{-2} + \Omega_\Lambda(1+z)^{-4}}}. \tag{1.53}$$

With this expression we are able to compute the comoving distance to an object only by observing its redshift, and assuming values for the Hubble constant and the energy density parameters. This set of fixed parameters, required to determine cosmic distances, is usually called a *cosmology*.

Angular diameter distance

The distance to an object can also be measured by observing the angle it subtends in the sky, and knowing its physical size. In a flat universe, the expression for this is

$$d_A \equiv \frac{l_p}{\theta} = a(t_e)\chi = \frac{\chi}{1+z}, \tag{1.54}$$

where l_p is the physical size of the object, and θ the observed angle.

In a static Universe, this would be equal to the physical distance to the object (since it comes from the geometrical definition of an angle). Nevertheless, since the scale factor varies, this is equal to the comoving distance to the object times the scale

factor at the time of emission of the observed radiation (i.e., the physical distance when the light was emitted).

For a non-flat Universe, the angular diameter distance is dependent on the curvature, and its expression is

$$\begin{aligned} d_{\mathrm{A}} &= \frac{a}{H_0\sqrt{\Omega_{\mathrm{k}}}}\sinh\left(\sqrt{\Omega_{\mathrm{k}}}H_0\chi\right) \quad \text{if} \quad \Omega_{\mathrm{k}} > 0 \\ d_{\mathrm{A}} &= \frac{a}{H_0\sqrt{|\Omega_{\mathrm{k}}|}}\sin\left(\sqrt{|\Omega_{\mathrm{k}}|}H_0\chi\right) \quad \text{if} \quad \Omega_{\mathrm{k}} < 0. \end{aligned} \tag{1.55}$$

Luminosity distance

Another method to determine the distance of an object is by comparison of its observed flux to its actual luminosity, if the latter is known. The relation between the observed flux F (at the time of observation t_0) and the luminosity of the object L is

$$F = \frac{L}{4\pi d_{\mathrm{L}}^2}, \tag{1.56}$$

where d_{L} is the *luminosity distance*. If the Universe was stationary, this would be equal to the comoving distance of the object, but given that the Universe expands, this is not the case.

The flux is the number of photons emitted by the source that cross a spherical shell per unit of (comoving) surface and time. Even if the shell is in comoving units, the wavelength of the photons will increase as the universe expands, and thus its energy will decrease by a factor of a (with a being the scale factor at the time of emission). Besides, the amount of photons that cross the shell in a unit of time will also decrease by a factor of a (since the distance they have to travel at light speed has increased by a factor of a^{-1}. Therefore, the relation between observed flux and luminosity will be:

$$F = \frac{La(t_{\mathrm{e}})^2}{4\pi\chi^2}, \tag{1.57}$$

with χ being the comoving distance to the object. Consequently, the luminosity distance is defined as

$$d_{\mathrm{L}} \equiv \frac{\chi}{a(t_{\mathrm{e}})} = (1+z)\chi. \tag{1.58}$$

Figure 1.5 shows the evolution versus redshift of these three distances, in logarithmic scale, for the cosmology defined in table 1.1. It is worth noting that d_{A} actually has a maximum value around $z \sim 1.5$, and then decreases again with redshift. This result may seem counter-intuitive, but it is simply due to the contraction of space at distances large enough, which increases the observed angle θ of an object (since space is smaller, an object of size l_{p} occupies a larger fraction of the sky, and thus subtends a larger angle).

Particle horizon and the observable Universe

Another concept related to distance is that of the limits of the observable Universe. According to general relativity, information can not travel faster than the speed of light c, and thus, an event at a certain point in space and time can only be perceived

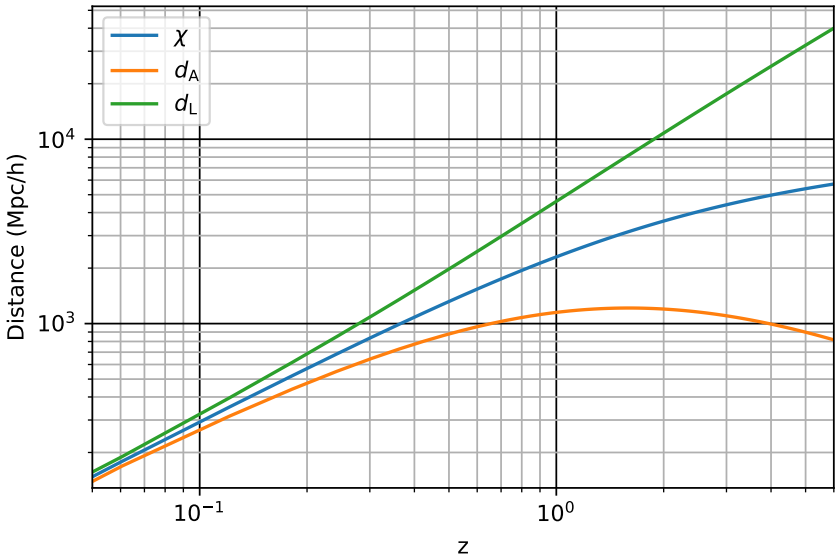

FIGURE 1.5: Comoving distance χ, angular diameter distance d_A and luminosity distance d_L versus redshift, for the cosmology specified in table 1.1.

in the regions of space where light has had time to travel to. Similarly, an observer can only perceive events from regions in space close enough so that light has had the time to arrive to said observer. These two space-time regions are called the future and past *lightcones* for a certain point in space-time, respectively; everything inside the lightcone is in *causal contact* with the observer. Figure 1.6 shows a schematic representation of a lightcone, for a three-dimensional space-time.

Given that the Universe has finite age, we can only observe the regions of the Universe that fall into our past lightcone, which is usually known as the observable Universe. At any given time, this volume corresponds to a sphere centred in us with a certain comoving radius, the *particle horizon*. This particle horizon is computed using eq. (1.51), but replacing t_e by zero (the start of the Universe), and t_0 by the time at which we want to determine it (for our current particle horizon, t_0 would remain the same).

This brings an interesting point that will be discussed in section 1.3.1: if a increases fast enough, the particle horizon can actually decrease with time. In order to do so, the Hubble flow at the particle horizon itself must be faster than the c. If that is the case, in spite of light always advancing at speed c, the Universe will expand even faster, which will mean that space will exit the particle horizon (and our observable Universe) as time advances.

The Hubble horizon

While the particle horizon represents a strict causal limit of how far can be we observe, another boundary of physical interest is the *Hubble horizon*, i.e., the surface at which the velocity of recession due to the Hubble flow surpasses the speed of light. This limit marks the point where the growth of structure is decoupled from physical processes, as will be discussed in section 2.2.2. It is simply defined as (in comoving units)

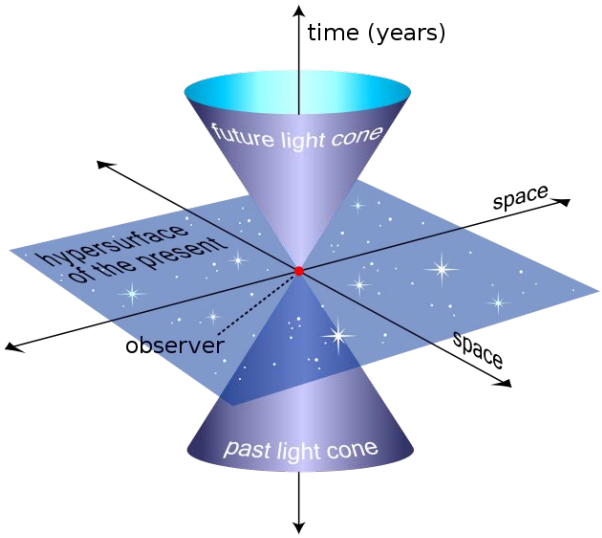

FIGURE 1.6: Schematic representation of a lightcone. Vertical axis represents time, horizontal surface space. Credit: MissMJ, Wikimedia Commons.

$$\chi_{\mathrm{H}}(t) = \frac{c}{aH(t)}, \tag{1.59}$$

and evolves differently from the particle horizon. In fact, for the current standard model, the former is always smaller (except the very brief inflationary period). This results in a sizeable portion of the observable Universe (the volume inside the particle horizon), drifting away from us at speeds larger than light. Figure 1.7 shows this result for the cosmology displayed in table 1.1.

Although the relative speed between two objects can not be greater than c, the geometrical expansion of space is not bound by this hard limit; thus, the evolution of the scale factor a may result in objects drifting away at speeds higher than c, if the distance between them is large enough. This does not violate general relativity, since this apparent relative motion is due to the expansion of space itself. In fact, if there is no relative motion between both objects, they will remain still in comoving space in spite of the expansion of space.

Nevertheless, the fact that such objects may be observed (as they are objects are inside the particle horizon) may also seem in contradiction with general relativity. However, what we actually observe is the radiation emitted by these objects long ago. If these objects were inside the Hubble horizon at the time of emission (even if they are receding superluminally at the time of observation), the observed photons have no information of its source moving away from the observer at speeds larger than c.

Besides, even if the observed objects are already outside of the Hubble volume when the radiation was emitted, the emitted photons can enter the Hubble horizon at some point in time. The relative velocity between the source and the Hubble horizon of the observer is always going to be smaller than between the source and the observer, so as long as the source doesn't move away from the expansion of our Hubble horizon at speeds larger than c, the emitted photons will eventually enter the Hubble volume. Once inside the Hubble volume, these photons will reach us in a

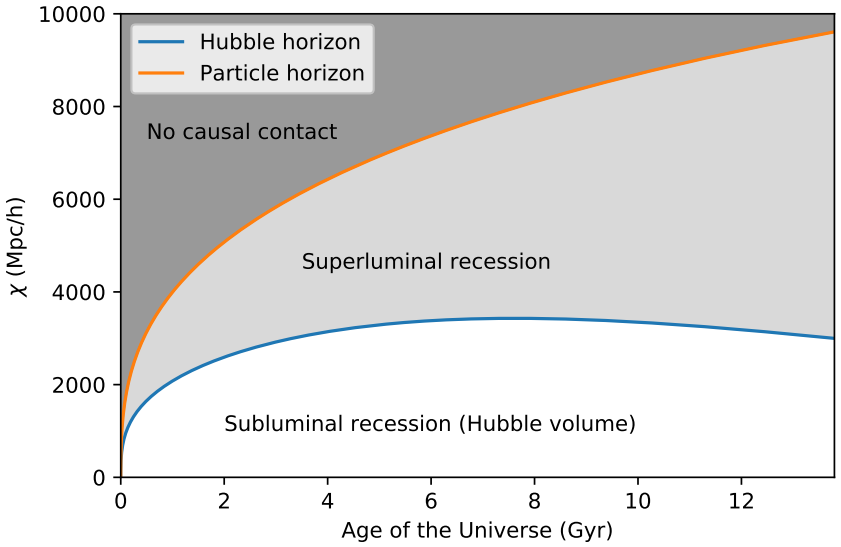

FIGURE 1.7: Evolution of the particle horizon and the Hubble horizon for the cosmology specified in table 1.1.

finite time, as c will always be larger than the speed of the Hubble flow in this region. The observed redshift of these photons will not be infinite either, as cosmological redshift is a product of the expansion of the universe along the trajectory of the photon, not the relative velocities between source and observer. (eq. (1.49)).

It is worth noting that, while the particle horizon is usually designated by that name in cosmological literature, the nomenclature for the Hubble horizon is far more diverse. The term *Hubble radius* is used interchangeably, and in some cases instead of referring to the limit of superluminal recession, authors refer to the volume enclosed by that limit as *Hubble volume* or *Hubble sphere*.

1.3 Brief history of the Universe

After having described the different components of the Universe, as well as its evolution according to the FLRW metric derived from general relativity, we can conclude this section by explaining in a very schematic way how this evolution is divided in different eras. This division is performed according to the predominant physical components at each era, and thus represent different regimes where the equations governing the Universe's evolution vary greatly.

1.3.1 Inflation (The very beginning)

According to the current paradigm of the standard model of cosmology, the first event right after the Big Bang was inflation. In a very short lapse of time ($\sim 10^{-32}$ s) the Universe expanded exponentially, which froze the quantum fluctuations that appeared in the original singularity, stretching them out to cosmic proportions. This rapid expansion, which implied an increase in a of several orders of magnitude (at least 10^{26}), also meant that the particle horizon decreased dramatically, since c was negligible compared to such an expansion rate.

This relatively simple model addresses some of the most important questions that do not seem to be explained with arbitrary initial conditions, namely:

- Existence of structure: At scales small enough so that the cosmological principle does not hold anymore, there are clear inhomogeneities (chapter 2), which eventually aggregate to form galaxies. The initial quantum fluctuations, expanded due to inflation, provide inhomogeneities in the initial density field that explain the formation of these structures.

- The horizon problem: The Universe at large scales is extremely homogeneous, even in regions far away enough to not be in causal contact yet (e.g., CMB regions separated by more than 60°). In principle, since these regions have never had any information of each other, there is no reason for them to be identical. However, in the inflationary model the particle horizon shrunk enormously during inflation, and resumed its growth once inflation ended; therefore, a conoving volume much larger than the current observable Universe was in causal contact before inflation, which explains the observed homogeneity.

The current standard model of inflation is the *slow-roll inflation*, where this rapid expansion is attributed to an scalar field that rolls down a potential, much more slowly than the actual expansion of the Universe (Linde, 1982; Albrecht and Steinhardt, 1982). Once this scalar field started to decay faster, inflation ended. With this fast decay, the cooling of the Universe due to the rapid expansion reverts, since the potential of the inflaton field is converted into particles. This process is usually known as *reheating*.

1.3.2 Radiation era (Almost the beginning-370 kyr)

After the end of inflation, the Universe was initially dominated by radiation, as can be seen in fig. 1.4. During this era, the Universe was extremely hot and dense compared to its actual state, and kept expanding, albeit at a much slower rate than inflation. In fact, by solving eq. (1.16) neglecting all terms but radiation density (eq. (1.32)), it can be shown that during this era $a \propto t^{1/2}$.

As the Universe slowly cooled down while expanding, several processes took place at a subatomic level. Some of the most important, in assumed chronological order, are:

- Baryogenesis: The quarks (elementary indivisible particles) formed from the decay of the inflaton field combined between themselves to form more complex particles. Baryons (neutrons and protons), which virtually account for all the observed, non-dark matter in the Universe, were formed at this stage.

- Decoupling of neutrinos: Neutrinos, extremely light particles that barely ever interact with matter or radiation, decoupled from other particles shortly after, and started freely travelling. This implies the existence of a cosmic neutrino background emitted before the CMB. Nevertheless, given how difficult it is to detect neutrinos, it is unlikely that this background signal will ever be detected in the foreseeable future.

- Nucleosynthesis: The protons and neutrons generated during baryogenesis combined to form nuclei more complex that ^{1}H, which consists only of one proton. This resulted in primordial abundances of ∼75% hydrogen, ∼25% helium, and trace amounts of lithium, which has been extensively verified with observations. All the elements other than hydrogen and helium (called *metals*

in cosmology) were generated much later, at the core of stars. To differentiate this primordial nucleosynthesis from subsequent ones, this phenomenon is usually referred to as Big Bang Nucleosynthesis (BBN).

The radiation era ends at the epoch of matter-radiation equality, where $\rho_{\mathrm{r}} = \rho_{\mathrm{m}}$, approximately at $z_{\mathrm{eq}} \sim 3600$ (around 50000 years in fig. 1.4, counting neutrinos as radiation).

1.3.3 Matter era (370 kyr-10 Gyr)

As matter became the main component of the Universe, the dynamics of its evolution changed; since radiation pressure was no longer dominant, baryonic matter could clump together due to the gravitational force, and form structures. Besides, expansion kept going at a faster rate (solving eq. (1.16) with only matter density from eq. (1.30) shows that $a \propto t^{2/3}$). The following events during this era led to the Universe we observe today:

- Recombination and the CMB: After the radiation era, photons were still coupled (i.e., in thermal equilibrium) with matter, which prevented them from travelling freely, and kept matter in an ionised state. When the Universe cooled down enough (at $z_{CMB} \sim 1100$), photons decoupled from matter, which caused two simultaneous events. First, photons started freely streaming, forming the CMB radiation we observe today. Second, the decoupled nuclei quickly bonded with electrons, turning into neutral atoms.
- The dark ages: Although the CMB had been released, hydrogen still had not clumped to form stars and galaxies. Besides, due to recombination hydrogen was in neutral state, and thus able to absorb radiation to excite its electron or be ionised. Combined, with the much higher gas density due to the small scale factor, it meant that right after the CMB the Universe was not only dark, but also mostly opaque.
- Reionisation: As the first stars and galaxies formed from gravitational collapse, the radiation emitted by them started ionising the surrounding medium, which rendered the Universe progressively transparent (given that ionised hydrogen can not absorb radiation). This was a gradual process, where ionisation started in bubbles around the first galaxies and ended up covering the whole space. The start of reionisation is believed to have started around $z \sim 20$, while its end is usually considered around $z \sim 6$.

After reionisation, at $z < 6$, the state of the Universe was, at first sight, very similar to what can be seen today: a transparent medium full of galaxies. Therefore, observations in the optical and near-infrared spectrum rarely go at $z > 6$, since from that point the further we go back in time the less transparent the Universe is, and the less galaxies there are to observe.

1.3.4 Dark energy era (10 Gyr-13.7 Gyr, i.e. today)

The matter era can be considered the bulk of the history of the Universe so far, but in "recent" times a new element has become the dominant component of the Universe; dark energy. Similarly to the radiation and matter era, the transition from the matter to the dark energy era is considered at the point where $\rho_m = \rho_\Lambda$, which happened at $z \sim 0.3$.

In contrast to the previous eras, during this dark energy phases that still lasts the Universe has not suffered any dramatic transformation in physical composition, such as baryogenesis or recombination. Therefore, the only major difference between the matter era and the dark energy is the accelerated expansion of the Universe, caused by the latter. In fact, solving eq. (1.16) while keeping only the Λ term from dark energy shows that $a \propto e^t$ in this era.

According to the ΛCDM model, such a exponential expansion will go on indefinitely, which implies the reduction of the observable Universe, which will eventually make unreachable any objects that are not gravitationally bound (i.e., anything farther away than the Milky Way and its surrounding galaxies in our case).

Figure 1.8 shows a diagram of the different stages of the evolution of the Universe, focused on the events after CMB (matter and dark energy eras).

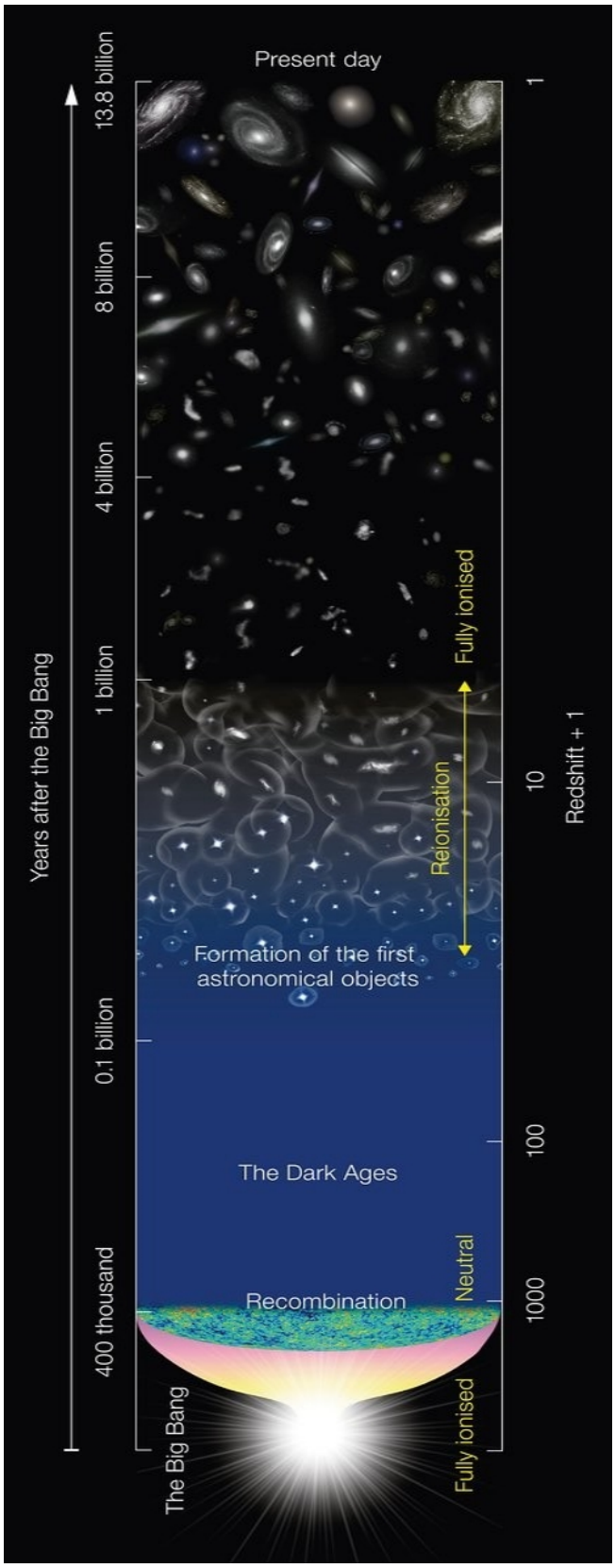

FIGURE 1.8: Cartoon of the evolution of the Universe, from the Big Bang to the present day. Credit: NAOJ.

Chapter 2

Large-scale structure

After explaining the basic theoretical foundations of cosmology in the previous chapter, here we will delve into the study of one of the most important features of the Universe from a cosmological standpoint: the large-scale structure.

The cosmological principle (i.e., the assumption that on scales large enough, the Universe is isotropic and homogeneous), is a core element of the great majority of cosmological studies carried out today. In fact, all the concepts laid out in section 1.2 (the scale factor and its evolution, distance measurements, etc.) rely completely on this principle.

Nevertheless, when we observe at scales small enough, it is clear that the cosmological principle is no longer valid. Due to the attractive effect of gravity, matter congregates in the densest areas, while the regions with smaller density progressively become more devoid of matter. This phenomenon results in a hierarchical, bottom-up formation of structure: baryonic matter clumps to form stars and other celestial bodies, which aggregate to form galaxies, and most galaxies are gravitationally bound to others in clusters. Galaxy clusters form even larger structures: filaments and nodes of a cosmic web.

This cosmic web is the largest structure that appears in the Universe; if averaged over scales large enough (roughly of the order of $\sim$100 Mpc/h, e.g., Gonçalves et al., 2018; Pandey and Sarkar, 2015), the cosmic web is indistinguishable and the Universe becomes completely homogeneous. Figure 2.1 shows this cosmic web, both mapped from observations (left), and generated in numerical simulations (right).

Dark matter is also hypothesised to form gravitationally bound structures known as *halos*. These structures are generally larger than galaxy scales, and may contain one or more galaxies inside. More compact dark matter structures are not expected in the ΛCDM paradigm, as dark matter is collisionless, and thus unable to accrete (i.e., lose energy and angular momentum due to collisions and friction during a gravitational collapse process). Regardless, these dark matter halos can be seen as the fundamental constituents of the cosmic web.

This ensemble of inhomogeneites arranged in a structured way is referred to as *large-scale structure* in cosmology, and it is an invaluable probe to determine the content of the universe, and to constrain different cosmological models as well. In order to do so, the properties of the large-scale structure need to be measured with well-defined statistics, and these statistics need to be related to the actual physical parameters of the Universe via theoretical models.

In this chapter, we will explain the most important statistics applied to the study of the large-scale structure: the two-point correlation function and its analogue in Fourier space, the power spectrum. Besides, we will finish with a small overview on a particular probe of large-scale structure of special interest for this thesis: the Lyα forest.

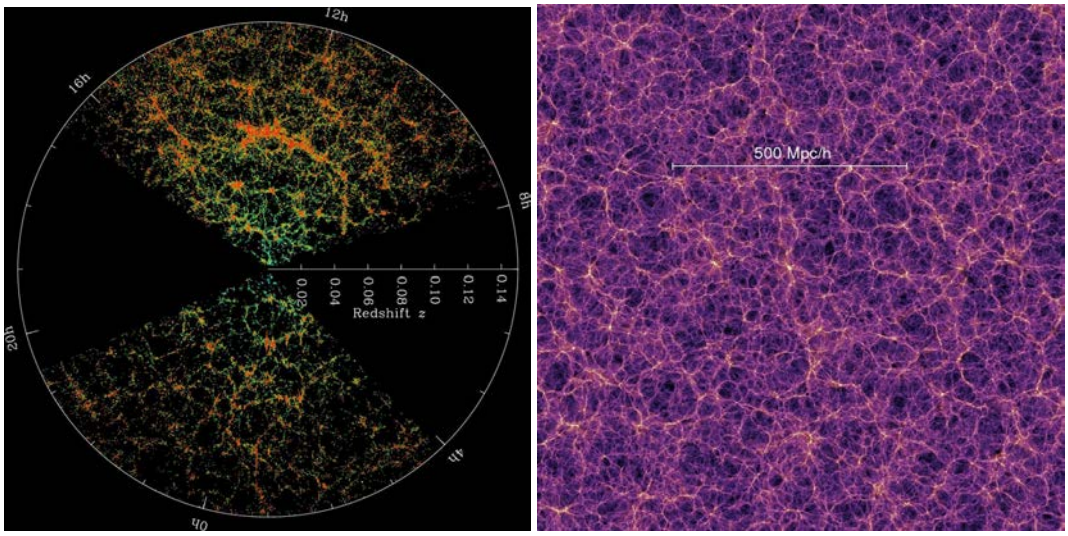

FIGURE 2.1: *Left panel:* Map constructed with the galaxies observed by the Sloan Digital Sky Survey (SDSS) in the near Universe. Credit: SDSS. *Right panel:* Snapshot at $z = 0$ of the dark matter distribution of the Millennium Simulation. Credit: Millennium Simulation Project.

2.1 Overdensities and the two-point correlation function

2.1.1 Overdensity

In order to quantify these inhomogeneities and their associated structure, it is more convenient to define an observable directly related to the variation of density relative to an average value, instead of working with absolute densities. The reasoning behind this choice is similar to the use of comoving distances: since the Universe has always been expanding, any absolute density is bound to monotonically decrease in time for a volume large enough. Therefore, a theoretical framework based on relative density variations will be valid at any point in time, without any need to scale.

This relative density variation is known as the *overdensity* (also called *density contrast*), and is defined as

$$\delta(\mathbf{x}, t) \equiv \frac{\rho(\mathbf{x}, t) - \langle \rho(t) \rangle}{\langle \rho(t) \rangle}, \tag{2.1}$$

where $\rho(\mathbf{x}, t)$ is the matter density at a certain point of space and time, and $\langle \rho(t) \rangle$ is the average matter density of the Universe at that given time.

If no index is added to δ, the overdensity is always assumed to be respective to the matter distribution (dominated by dark matter). Regarding the other components of the Universe discussed in section 1.2.4, radiation only plays an important role well before the CMB, and it is completely negligible in the observable Universe (see fig. 1.4), while dark energy is assumed homogeneous (and only relevant at late times). Thus, the study of the large-scale structure of the Universe is the study of its matter distribution.

2.1.2 Bias

The overdensity δ, as defined in 2.1, is an expression of the inhomogeneities in the dark matter distribution. Dark matter, however, does not interact with radiation,

and thus can not be directly observed. What we observe instead is radiation coming from baryonic matter (e.g., galaxies, quasars, intergalactic medium (IGM), etc.), or the effects of matter on radiation (weak lensing). When referring to the dark matter distribution, these observables are called *tracers* of dark matter.

Consequently, the observed δ_t for a given tracer t will be biased with respect to the actual dark matter distribution. This bias b_t is always modelled as a multiplicative factor of δ:

$$\delta_t(\mathbf{x}, t) = b_t \delta(\mathbf{x}, t). \tag{2.2}$$

The most simple functional form of this bias $b_t$is to assume that it is constant, however, this can only be applied as a first-order approximation. An accurate modelling of the bias will result in dependencies not only in redshift and the mass of the respective halo the object is contained in (the *classical* bias parameters), but also the masses of surrounding halos, their merger history, or even other physical properties such as the halo spin. These contributions of the environmental properties to the bias of a certain are known as *assembly bias*.

In this work, we will only use simplistic assumptions of the two tracers that will be cross-correlated (the Lyα forest and the diffuse Lyα emission), partially because the current constraints on the values of these biases is very limited, and also because a detailed modelling of the bias is out of the scope of this thesis. For a comprehensive review of the literature on galaxy (and other tracers) bias, see Desjacques, Jeong, and Schmidt, 2018.

2.1.3 The two-point correlation function

Out of all the statistics that can be estimated from the observed distribution of a tracer, the most used in cosmology (at least in real space) is, by far, the two-point correlation function (2PCF). For a scalar field dependent on space (such as the overdensity $\delta(\mathbf{x})$), the 2PCF for a certain distance vector $\mathbf{r}$ is defined as the average value of the product between $\delta(\mathbf{x})$ and the field at the points at a certain distance $\mathbf{r}$, $\delta(\mathbf{x}+\mathbf{r})$. Mathematically, this is expressed as

$$\xi(\mathbf{r}) \equiv \frac{1}{V}\int_V d\mathbf{x}\, \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r}) = \langle \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle. \tag{2.3}$$

Where V is the volume of integration. In a purely mathematical scenario, ideally $V = \mathbb{R}^3$, however, when working with observational data or simulations the actual volume where $\delta(\mathbf{x})$ is defined will be finite, which will introduce artificial variability in the calculation of $\xi(\mathbf{r})$. Usually, if the volume of integration is large enough (approximately 10 times larger than $|\mathbf{r}|$ in the smallest dimension), these effects are fairly negligible compared to observational error and cosmic variance.

Given the cosmological principle, if this volume of integration V is large enough, the 2PCF will be isotropic, since at any point $\mathbf{x}$ the Universe on average must look the same regardless of the direction of the $\mathbf{r}$ vector. Therefore, the 2PCF will depend simply on the scalar variable r, the modulus of $\mathbf{r}$, which can be expressed as

$$\xi(r) \equiv \langle \delta(\mathbf{x})\delta(\mathbf{x}+\mathbf{r})\rangle \quad \forall |\mathbf{r}| = r. \tag{2.4}$$

This theoretical 2PCF is defined for the total matter distribution, dominated by dark matter; any actual observations will be based on tracers with a certain bias, as expressed in eq. (2.2). Therefore, for any tracers $t1$ and $t2$, the measured correlation will be related to the real matter 2PCF as

$$\xi_{t1\,t2}(r) = b_{t1} b_{t2} \xi(r), \tag{2.5}$$

where b_{t1} and b_{t2} are the respective bias of the tracers. If $t1$ and $t2$ are the same sample of the same tracer, the effective bias of the 2PCF will simply be b_{t1}^2, and in such case the measured 2PCF is usually referred to as *autocorrelation*. If the tracers being correlated are different, then it is known as *cross-correlation*.

When this 2PCF is estimated from observational data, the tracers often are not scalar fields, but discrete counts in redshift space (e.g., galaxies mapped by a survey). For the autocorrelation of galaxies (or other discrete counts), the 2PCF has a far more intuitive meaning: it is the excess in probability of finding a galaxy at a distance r from another galaxy with respect to a Poisson distribution (where the probability of finding a galaxy at a certain point is completely independent from its environment).

Therefore, a positive (negative) value of the 2PCF at a certain distance r means that the probability of finding two galaxies as such distance is higher (lower) than in a fully random distribution. Some simple properties can be deduced from this interpretation of the 2PCF. First, for distances large enough $\xi(r)$ must tend to zero, since above the scale of homogeneity the distribution should be indistinguishable from a random distribution. And second, the cumulative integral $\int \xi(r) dr$ also must tend to zero at the same scales, given that the cumulative probability becomes that of a random distribution as well.

In fig. 2.2 an example of matter 2PCF is displayed, both its theoretical prediction (solid line) and the values measured from a simulation (blue points). This was a validation test for the correlation code developed for this work (section 5.2). For a complete review on the 2PCF and its estimators, as well as other statistics applied to the study of large-scale structure, we refer the reader to Bernardeau et al., 2001, chapter 3.

2.2 The power spectrum

2.2.1 Definition of the power spectrum

From a mathematical standpoint, the evolution of the structure observed in the Universe according to its physical parameters is better described in Fourier space. First, we define the Fourier transform of a certain scalar field $f(\mathbf{x})$ defined in $\mathbb{R}^3$ as

$$\mathcal{F}[f(\mathbf{x})] \equiv \int_{\mathbb{R}^3} d^3\mathbf{x}\, f(\mathbf{x}) e^{-i\mathbf{kx}} = \widetilde{f}(\mathbf{k}), \tag{2.6}$$

as well as its inverse Fourier transform

$$\mathcal{F}^{-1}[\widetilde{f}(\mathbf{k})] \equiv \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} d^3\mathbf{k}\, \widetilde{f}(\mathbf{k}) e^{i\mathbf{kx}} = f(\mathbf{x}). \tag{2.7}$$

The $(2\pi)^{-3}$ factor in the inverse Fourier transform comes from the normalisation criteria applied. Other normalisations of the transform, which would yield different coefficients in $\mathcal{F}$ and $\mathcal{F}^{-1}$ are entirely possible, but this one is arguably the most used convention is cosmology.

Another necessary definition for this explanation is that of the convolution of two scalar fields, $f(\mathbf{x})$ and $g(\mathbf{x})$, which is expressed as

$$(f * g)(\mathbf{x}) \equiv \int_{\mathbb{R}^3} d^3\mathbf{x}' f(\mathbf{x} - \mathbf{x}') g(\mathbf{x}'). \tag{2.8}$$

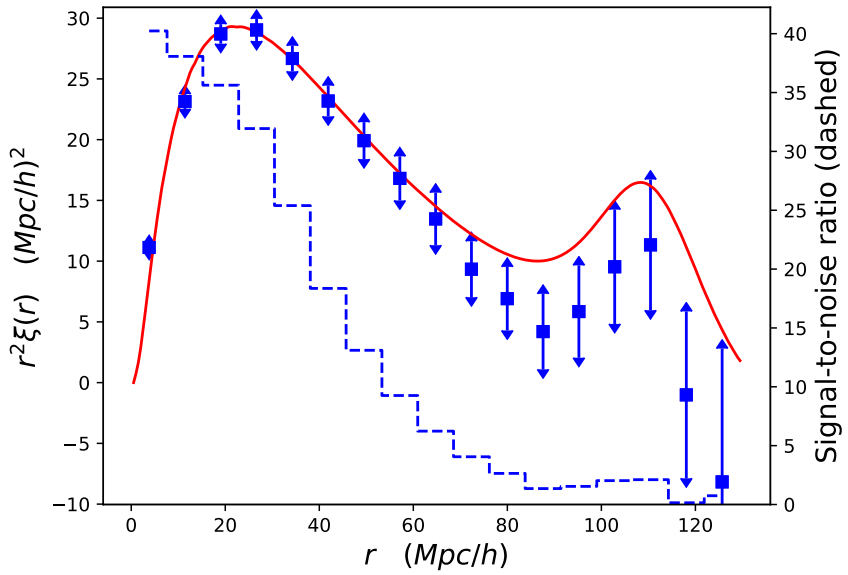

FIGURE 2.2: Example of 2PCF used to validate the correlation code developed for this thesis. The blue points represent the 2PCF measured by the code in the MICE Grand Challenge Simulation (Fosalba et al., 2013), the solid red line is the theoretical prediction using the cosmology of the simulation, and the dashed blue line is SNR of each bin (right axis). The increased error with distance is due to the limited sample volume.

This operator has an interesting property in Fourier space, named the *convolution theorem*, which states that the Fourier transform of the convolution of two functions is equal to the products of the Fourier transform of each individual function,

$$\mathcal{F}[(f * g)(\mathbf{x})] = (2\pi)^3 \mathcal{F}[f(\mathbf{x})]\mathcal{F}[g(\mathbf{x})]. \tag{2.9}$$

The $(2\pi)^3$ factor in this expression comes from the chosen normalisation of the Fourier transform; different normalisations would result in different factors on the left-hand side of eq. (2.9).

In order to demonstrate this convolution theorem, we write eq. (2.8), but replacing the function $f(\mathbf{x})$ inside the integral by $\mathcal{F}^{-1}[\widetilde{f}(\mathbf{k})]$, while keeping the $x - x'$ term,

$$(f * g)(\mathbf{x}) = \int_{\mathbb{R}^3} d^3\mathbf{k} \widetilde{f}(\mathbf{k}) e^{i\mathbf{k}(\mathbf{x}-\mathbf{x}')} \int_{\mathbb{R}^3} d^3\mathbf{x}' g(\mathbf{x}'), \tag{2.10}$$

and rearranging the terms inside the integral, we find

$$(f * g)(\mathbf{x}) = \int_{\mathbb{R}^3} d^3\mathbf{k} \widetilde{f}(\mathbf{k}) \left[\int_{\mathbb{R}^3} d^3\mathbf{x}' g(\mathbf{x}') e^{-i\mathbf{k}\mathbf{x}'} \right] e^{-i\mathbf{k}\mathbf{x}} = \int_{\mathbb{R}^3} d^3\mathbf{k} \widetilde{f}(\mathbf{k}) \widetilde{g}(\mathbf{k}) e^{-i\mathbf{k}\mathbf{x}}. \tag{2.11}$$

The left-hand side is the inverse Fourier transform of the product of $\widetilde{f}(\mathbf{k})$ and $\widetilde{g}(\mathbf{k})$, i.e.,

$$(f * g)(\mathbf{x}) = (2\pi)^3 \mathcal{F}^{-1}[\widetilde{f}(\mathbf{k})\widetilde{g}(\mathbf{k})], \tag{2.12}$$

so simply by applying the Fourier transform to both sides of eq. (2.12), we recover eq. (2.9).

Having defined the convolution operator and its properties in Fourier space, it is interesting to note that the 2PCF, as defined in eq. (2.3), is the convolution of δ with itself at the point $-\mathbf{r}$ (i.e., $\xi(\mathbf{r}) = (\delta * \delta)(-\mathbf{r})$). If we take isotropy into account, and consider r a scalar variable, then we have $\xi(r) = (\delta * \delta)(r)$. Therefore, the 2PCF becomes a simple product of two functions in Fourier space. Since it is the main statistic applied to the analysis of the structure of the Universe, it is now clear why would it be more convenient to develop its associated physical theory in this space. In order to do so, we define the power spectrum of the matter distribution, $\mathrm{P}(\mathbf{k})$, as

$$P(\mathbf{k}) \equiv |\widetilde{\delta}(\mathbf{k})|^2, \tag{2.13}$$

where $\widetilde{\delta}(\mathbf{k})$ is the Fourier transform of the overdensity field $\delta(\mathbf{x})$. It is worth noting that, according to this definition, $P(\mathbf{k})$ has units of (length)^3, since $\delta(\mathbf{x})$ is dimensionless (eq. (2.1)), and $\widetilde{\delta}(\mathbf{k})$ has units of (length)^3, considering that in our definition Fourier transform is an integral over $\mathbb{R}^3$, and $\mathbf{k}$ has units of (length)$^{-3}$.

With this definition, the 2PCF is simply the inverse Fourier transform of the power spectrum, i.e.,

$$\xi(r) = \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} d^3\mathbf{k} P(\mathbf{k}) e^{i\mathbf{kx}} = \frac{1}{2\pi^2} \int_0^\infty dk\, k^2 P(k) e^{ikx}. \tag{2.14}$$

Here, in the left-side of the equation we have assumed isotropy (due again to the cosmological principle), so $P(k)$ only depends on the modulus of $\mathbf{k}$ ($k = |\mathbf{k}|$), and integrated over the solid angle of a sphere in spherical coordinates (which is equivalent to multiplying by 4π for a isotropic field). Therefore, the power spectrum is usually expressed as a function of k. The values of this variable k are usually known as *modes*; for a certain mode k, a higher value of $P(k)$ means a larger degree of inhomogeneity at these scales (i.e., more dispersion around zero of the values of $\delta(\mathbf{x})$).

So far, we have discussed the power spectrum of the matter distribution, dominated by the underlying dark matter. In the same fashion as the 2PCF, the power spectrum of a certain tracer t will be biased with respect to the matter power spectrum. Given that the Fourier transform is linear, the relationship between the matter power spectrum $P(k)$ and the power spectrum of the tracer $P_t(k)$ can be written as

$$P_t(k) = b_t^2 P(k), \tag{2.15}$$

where b_t is exactly the same bias for the tracer as described in eq. (2.2) for real space.

2.2.2 Theoretical model of the power spectrum

The theoretical modelling of the power spectrum and its evolution requires to consider the wildly different physical properties of the Universe in the different eras it goes through, as explained in section 1.3, as well as the evolution of all the modes of $P(k)$, which may seem like a daunting task at first.

However, a mostly analytical expression for the power spectrum can be obtained in the linear regime approximation ($\delta << 1$). Under this assumption, the equations governing the evolution of the modes of $P(k)$ can be expanded using perturbation theory, and all the non-linear terms (higher than first order) are neglected. An expression of the power spectrum as a function of k and z in the linear regime is

$$P(k,z) = P_0(k)T(k)^2D(k,z)^2. \tag{2.16}$$

- $P_0(k)$ is the primordial power spectrum, given by the conditions right after inflation.
- $T(k)$ is the transfer function, which models how the modes go through the radiation era until the matter era.
- $D(k,z)$ is the growth factor, which dictates how the modes develop since the beginning of the matter era until the present day.

The reason why the transfer function and the growth factor are squared in eq. (2.16) is that they are defined to describe the power spectrum of the gravitational potential, not the matter fluctuations.

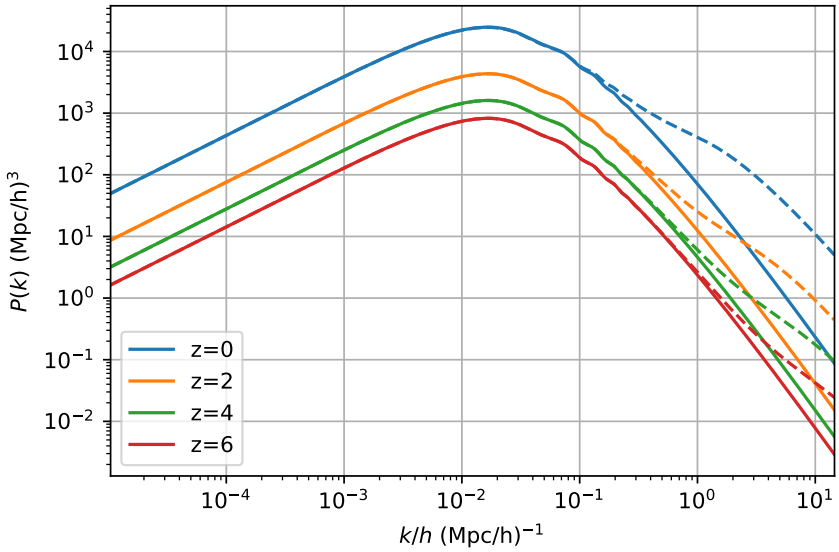

FIGURE 2.3: Linear power spectrum (solid line) and its non-linear counterpart (dashed line), for different redshifts, computed using CAMB for the cosmology in table 1.1.

The primordial power spectrum

The current inflationary paradigm considers that right at the end of inflation, the only inhomogeneites in the Universe were the quantum fluctuations stretched to cosmic scale. If these fluctuations come from a single non-interacting field, they must follow a Gaussian distribution. Consequently, the primordial power spectrum right after inflation must be that of a Gaussian distribution, which adopts the form of a power law (Guth and Pi, 1982) such as

$$P_0(k) = Ak^{n_s}. \tag{2.17}$$

Here, A is a normalisation constant that fixes the amplitude of the power spectrum, and n_s, the exponent of the power law, is the *spectral index*. If $n_s = 1$, we usually refer to a *scale invariant* power spectrum (referring to the fact that, in a logarithmic scale of base k, $P_0(k)$ would be displayed as a constant).

These two variables are also cosmological parameters of the ΛCDM model; in the case of of n_s, it is given directly, while A is usually computed by adjusting the normalisation of the power spectrum to a certain observable. By convention, the statistic used for this purpose most of the time is σ_8, i.e., the standard deviation of the overdensity field in a sphere of radius 8 Mpc/h at $z = 0$. In table 1.1, the values of n_s and σ_8 for Planck Collaboration et al., 2016 are displayed.

The calculation of σ_8 for both real and Fourier space can be expressed as

$$\sigma_8^2 \equiv 4\pi \int_0^\infty dr\, W_r(r) r^2 \delta(r)^2 = \frac{1}{2\pi^2} \int_0^\infty dk\, W_k(k) k^2 P(k), \tag{2.18}$$

where $W_r(r)$ and $W_k(k)$ are the window functions that define the integration range in real and Fourier space, respectively. For the particular case of σ_8, these are top-hat functions of radius 8 Mpc/h in real space and 0.125 $(\mathrm{Mpc/h})^{-1}$ in Fourier space. Therefore, with an observed value of σ_8, using this expression the normalisation of the power spectrum given by the factor A can be computed, assuming that $P(k, z)$ follows a functional form as in eq. (2.16).

The transfer function

Once inflation ends, the primordial power spectrum generated by the Gaussian quantum fluctuations starts to evolve according to the physical properties of the Universe at that time: an extremely hot, radiation-dominated environment that progressively expands and cools down. Here we will shortly discuss the meaning and approximate behaviour of the transfer function that describes the evolution of the power spectrum during this era, as a derivation of the function is out of the scope of this thesis. For a comprehensive derivation of the transfer function and the growth factor, we refer the reader to Dodelson and Schmidt, 2020, chapter 7.

Qualitatively, the evolution of overdensity with time can be understood as

$$\frac{d^2\delta}{dt^2} = (\text{Gravity - Pressure}) \cdot \delta, \tag{2.19}$$

i.e., since it is attractive and dependant on mass, gravity accelerates the growth of overdensities, while pressure decelerates it. If pressure dominates, the evolution of δ will become oscillatory. In this expression we have just added abstract terms to represent the effects of each force, without including an actual functional dependence on physical parameters.

In the radiation-dominated era, radiation pressure is the dominant contributor to the pressure term in eq. (2.19). Since dark matter does not interact with electromagnetic radiation, dark matter fluctuations still grow, but a slower rate than when matter dominates in the next era. Since radiation is the dominant energy component in the early Universe, it affects significantly the gravitational potential by homogenising it. As radiation is massless, and thus not attracted by gravity in the non-relativistic sense, it will not cluster in the lowest potential regions, generating even lower potentials (which is the basis for structure formation).

Given that the increase of potential is partially suppressed by radiation, this results in a logarithmic growth of the dark matter overdensity,

$$\delta \propto \ln a. \tag{2.20}$$

On the other hand, baryonic matter is coupled to radiation, and its pressure overcomes gravity when it is the dominant element of the Universe. Therefore, baryonic

fluctuations exhibit an oscillatory behaviour, according to eq. (2.19), while being progressively suppressed. Since most of the matter content is dark matter, these oscillations only show as small wiggles in the net power spectrum (highlighted in the transfer function displayed in fig. 2.4). Besides, when measuring the 2PCF in real/redshift space, a peak in this statistic around ∼110 Mpc/h (see fig. 2.2) appears for this same reason: it is the comoving length of the radiation pressure waves just before decoupling from baryonic matter (see section 2.3).

Nevertheless, radiation (or any other physical phenomenon) only affects modes inside the Hubble horizon (i.e., $k > k_{\mathrm{H}}$); larger modes are simply out of causal contact (as information can not travel fast enough to affect them), and thus are considered to be *frozen*. These modes, however, do still evolve, but isolated from any kind of causal phenomenon. It can be seen as a purely kinematic evolution, where modes evolve "on their own", as opposed of a dynamical evolution inside the Hubble horizon, where pressure and gravity influence their evolution. When discussing the evolution of power spectrum modes, the term *horizon* always refers to the Hubble radius.

A strict derivation of the behaviour of modes outside of the horizon would require a relativistic treatment of perturbation theory, and the standard Newtonian derivation is already out of the scope of this thesis. Nevertheless, the approximate evolution of super-horizon modes can be deduced from a very simple perturbative approach. First, let us assume the first Friedmann equation (eq. (1.16)), for a completely flat Universe ($k = 0$), and the cosmological constant Λ included in the energy density term (eq. (1.33)),

$$H(t)^2 = \frac{8\pi G}{3}\rho_0(t), \tag{2.21}$$

where ρ_0 would be the average energy density of the Universe at time t. If a small perturbation in density is introduced while keeping the same expansion rate, a non-negligible curvature will appear, i.e.,

$$H(t)^2 = \frac{8\pi G}{3}\rho_1(t) - \frac{k}{a(t)^2}. \tag{2.22}$$

Here, $\rho_1(t)$ is the perturbed value of the average density, and k is the curvature parameter, not the mode wavenumber. Recalling the definition of overdensity (eq. (2.1)), the fluctuation δ due to this small perturbation can be written as a function of the scale factor and the energy density as

$$\delta(t) = \frac{3k}{8\pi G a(t)^2 \rho_0(t)} \propto a(t)^{-2}\rho_0(t)^{-1}. \tag{2.23}$$

This expression comes from replacing in (eq. (2.1)) ρ_0 and ρ_1 on the denominator, using the values obtained from eq. (2.21) and eq. (2.22) respectively. Therefore, if we know the evolution of the average density and the scale factor at a given time, the super-horizon growth of δ can be estimated. For the radiation era, we know that $\rho \propto a^{-4}$ (eq. (1.32)), so the growth of super-horizon fluctuations as a function of the scale factor can be expressed as

$$\delta(a) \propto a^2. \tag{2.24}$$

Having defined the evolution of the fluctuations on both sub-horizon and super-horizon modes, we can finally describe the behaviour of the transfer function $T(k)$ (eq. (2.16)). By convention, this function is normalised so at the largest scales, $T(k) =$

1 (well outside the horizon at the epoch of equality, where the radiation era ends). Therefore, since all super-horizon modes evolve similarly, $T(k) \sim 1$ for these modes. On the other hand, the smaller-scale modes that entered the horizon in the radiation era will have $T(k) < 1$, since the modes tend to grow logarithmically, but the normalisation introduces an a^{-2} factor from the super-horizon growth.

The smaller the mode (larger k), the sooner it will enter the Hubble horizon, and stop growing quadratically. Consequently, sub-horizon modes with larger k will have a smaller value of $T(k)$, as its growth was slowed down by radiation earlier, while its super-horizon counterparts kept growing. The transfer function will then show the following behaviour

$$T(k) \propto \begin{cases} 1 & \text{if} \quad k \ll k_{\mathrm{H\,eq}} \\ k^{-2} \ln k & \text{if} \quad k \gg k_{\mathrm{H\,eq}} \end{cases} \tag{2.25}$$

with $k_{\mathrm{H\,eq}}$ being the wavenumber of the Hubble horizon at the epoch of equality. In practice, there is not a clear cut between the super-horizon and sub-horizon modes, and numerical codes are needed to provide an exact solution, since it requires to solve both the Einstein equations determining the evolution of matter and gravitational potential, as well as the Boltzmann equations that describe the evolution of the coupled photon fluid.

However, eq. (2.25) already illustrates the most important feature: the turnover of the transfer function around $k_{\mathrm{H\,eq}}$. This can be seen in both the transfer function displayed in fig. 2.4, as well as the power spectra in fig. 2.3. The primordial power spectrum is a power law very close to a linear slope (which should be displayed as linear in a logarithmic plot, regardless of the exponent); the characteristic form of the observed power spectrum, with a maximum around k_{H} at z_{eq}, comes from the turnover in the transfer function.

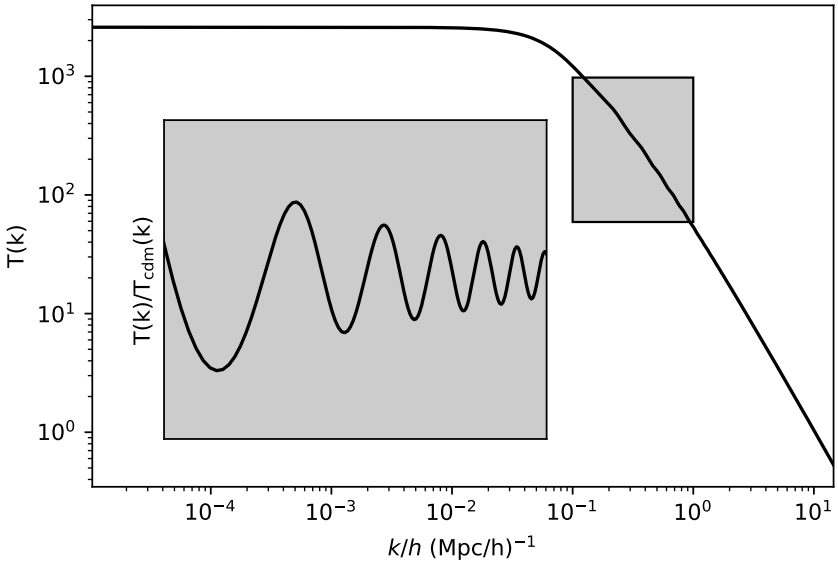

FIGURE 2.4: Transfer function with an arbitrary normalisation, as defined in eq. (2.16), for the cosmology in table 1.1, computed in CAMB. The shaded area is the regime where the baryonic wiggles are more noticeable; in the zoomed panel the ratio between $T(k)$ and $T_{\mathrm{cdm}}(k)$ is shown for this shaded region.

The growth factor and non-linearities

Once matter becomes the dominant element of the Universe, and radiation decouples some time after, describing the evolution of the power spectrum becomes a less daunting task. On scales large enough (of the order of Mpc/h), baryonic effects such as pressure or friction become negligible, and the matter distribution can be considered a perfect fluid. Besides, density fluctuations and velocities are small to allow not only a purely Newtonian approach, but also a solution via linear perturbation theory. The derivation shown in this subsection is extracted from Peebles, 1980.

First, since matter is to be considered a perfect fluid, the differential equations that determine its evolution need to be laid out. We will do so in a form already adapted for the FRLW metric, following the conventions explained in chapter 1. First, the continuity equation, which is akin to the condition of mass conservation,

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla(\mathbf{v} + \delta\mathbf{v}) = 0. \tag{2.26}$$

Second, the Euler equation, which expresses momentum conservation by establishing a relation between the net variation of momentum and the gravitational potential ϕ (neglecting pressure gradients),

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{1}{a}(\mathbf{v}\cdot\nabla) + \frac{\dot{a}}{a}\mathbf{v} = -\frac{1}{a}\nabla\phi. \tag{2.27}$$

And finally, the Poisson equation, which relates the gravitational potential to the matter content,

$$\nabla^2\phi = \frac{3}{2}H_0^2\Omega_{\mathrm{m}}a^{-1}. \tag{2.28}$$

In these equations, $\mathbf{v}$ represents the peculiar velocity of the matter fluid at the point where the equation is evaluated, i.e., the velocity minus the Hubble flow (in comoving space, $\mathbf{v}$ is actually the net velocity of the fluid).

We can apply the linear approximation to simplify eq. (2.26) and eq. (2.27); in order to do so, all the terms that are higher than first order in δ and $\mathbf{v}$ are neglected (e.g., δ^2, $\mathbf{v}^2$, $\delta\mathbf{v}$). The continuity equation is then

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\nabla\mathbf{v} = 0, \tag{2.29}$$

and moving the left-hand term to the right-hand side, the linear Euler equation is

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\dot{a}}{a}\mathbf{v} + \frac{1}{a}\nabla\phi = 0. \tag{2.30}$$

With these lineal approximations, a single differential equation for δ can be obtained with the following steps:

1. Apply time derivative to eq. (2.29)
2. Multiply eq. (2.30) by $a^{-1}\nabla$.
3. Subtract the second equation to the first; $\mathbf{v}$ should not appear in the resulting equation.
4. In the resulting equation, replace $\nabla^2\phi$ using eq. (2.28).

The result is a second-order differential equation where, for a given cosmology, the only variable is δ

$$\frac{\partial^2\delta}{\partial t^2} + 2\frac{\dot{a}}{a}\frac{\partial\delta}{\partial t} - \frac{3}{2}\Omega_{\rm m}a^{-3}H_0^2 = 0 \tag{2.31}$$

Considering that the overdensity is a function of both time and space ($\delta(\mathbf{x}, t)$), and that this equation only places constrains on time evolution, a solution for δ can adopt the following form:

$$\delta(\mathbf{x}, t) = A(\mathbf{x})D_+(t) + B(\mathbf{x})D_-(t), \tag{2.32}$$

where $A(\mathbf{x})$ and $B(\mathbf{x})$ are arbitrary spatial functions, and $D_+(t)$, $D_-(t)$ temporal functions whose value increases and decreases with time, respectively (*growing* and *decaying modes*). Given that we are evaluating the temporal evolution of δ, we can neglect the spatial part; in the linear regime the growth of fluctuations will be the same everywhere. Besides, since gravity is an attractive force (and the only force considered in eq. (2.28) is gravity indeed), the decaying mode $D_-(t)$ can be ruled out as a non-physical solution. The growing mode $D_+(t)$ (henceforth $D(t)$) is the growth factor we are looking for. Knowing this, eq. (2.32) can be rewritten only for $D(t)$,

$$\frac{\partial^2 D(t)}{\partial t^2} + 2H(z)\frac{\partial D(t)}{\partial t} - \frac{3}{2}\Omega_{\rm m}H_0^2(1+z)^3 D(t) = 0. \tag{2.33}$$

Where we have expressed the equation as a function of the Hubble parameter and redshift, to follow the conventional notation. For matter domination (all other energy densities negligible), we find that, since $a \propto t^{3/2}$, $D(t) \propto a$. A more general solution can be provided in integral form for a flat Universe with a cosmological constant, which corresponds to the ΛCDM model at late times. This solution, expressed as a function of redshift (which is the most common form for the linear growth factor), is

$$D(z) = \frac{H(z)}{H_0}\left[\int_0^\infty \frac{dz'(1+z')}{H(z')^3}\right]^{-1} \int_z^\infty \frac{dz'(1+z')}{H(z')^3}. \tag{2.34}$$

This linear growth factor only depends on redshift (time), and not space (or modes k); as can be seen in fig. 2.3, for different redshifts the linear power spectrum only shows an offset in the total amplitude, while conserving the same shape.

Nevertheless, this is only accurate in the linear regime (i.e., when δ is of the order of few at most, and $v \ll c$); when linear theory breaks down, each mode evolves differently. In fig. 2.3 this effect is clear, since the larger the k (smaller scales), the more noticeable is the power gain with time. The origin of this behaviour is gravitational collapse, and thus, when discussing non-linearities and non-linear regime in the power spectrum, it refers to the effects of gravity on small scales.

2.3 Baryon Acoustic Oscillations

As explained in section 2.2.2, during the radiation-dominated era, in the early Universe, photons were coupled to baryonic matter, and radiation pressure was high enough to overcome the attractive effect of gravity. This generated oscillations of the baryon-photon fluid, which can be seen as wiggles in the transfer function and the power spectrum (e.g., fig. 2.4).

These oscillations, however, translate to real space as longitudinal pressure waves that propagate with time, akin to sound waves in our atmosphere. When matter and radiation decoupled, baryons where no longer subjected to radiation pressure, which froze these acoustic oscillations. Consequently, this results in an imprint in the baryonic distribution: the fronts of these waves right before decoupling (i.e., the maximum distance travelled by the waves during the radiation era) will have larger baryon densities than average. This feature in the matter distribution is referred to as *baryon acoustic oscillations* (BAO), and the length of these oscillations is commonly known as the *sound horizon*.

While not directly observable by eye, BAO can be detected in clustering statistics such as the 2PCF (e.g., the peak in fig. 2.2 around ∼110 Mpc/h is indeed the BAO signal in the simulation). Initially, BAO were hypothesised to be present in the CMB power spectrum (Peebles and Yu, 1970), and also first observed there (Miller et al., 1999). However, the imprint of BAO should also be observable at later times in galaxy distributions, and eventually was detected in the 2PCF computed from large spectroscopic samples (Eisenstein et al., 2005; Cole et al., 2005).

The scientific value of the BAO observation is that they function as a standard ruler: since these are frozen pressure waves at scales too large for gravitational interaction, they do not collapse and simply expand at the same rate as the scale factor a, which makes their size constant in comoving space. BAO in the line-of-sight direction can then be used to precisely determine the Hubble parameter at a certain redshift, while transverse BAO yield a measurement of the angular diameter distance (Seo and Eisenstein, 2003).

These two observables combined allow to constrain the expansion history of the Universe, and thus constitute an independent probe to evaluate dark energy and its equation of state. Consequently, ever since its first detections BAO have been at the core of several observational programs, to the point of being one of the main goals of the spectroscopic surveys simulated in this work (see section 3.3, section 3.4).

2.4 The Lyα forest

The theoretical models of large-scale structure are mostly based on the underlying dark matter distribution, which is always the main contributor to the gravitational potential, while all observational data we can retrieve is from tracers that follow this dark matter distribution. While most of these tracers are directly observed thanks to the radiation they emit in different wavelengths (e.g., galaxies, quasars, hot gas), in this section we will discuss a tracer of special interest for this work, observed indirectly thanks to absorption lines in spectra: the Lyα forest.

2.4.1 Introduction to the Lyα forest

Most of the baryons in the Universe do not lie in galaxies (either in stars or as interstellar gas), but in the intergalactic medium (IGM) in the form of gas, clustered around the underlying dark matter web, but undetectable by direct means. Given its density, much lower than the galactic medium, any thermal or line emission that may come from the IGM is completely negligible compared to the observed galaxies and quasars, where the vast majority of optical emission is concentrated in stars and accretion disks.

The composition of these IGM clouds is by ∼75% hydrogen and ∼25% helium, following the BBN proportions. Most of this hydrogen nowadays is completely

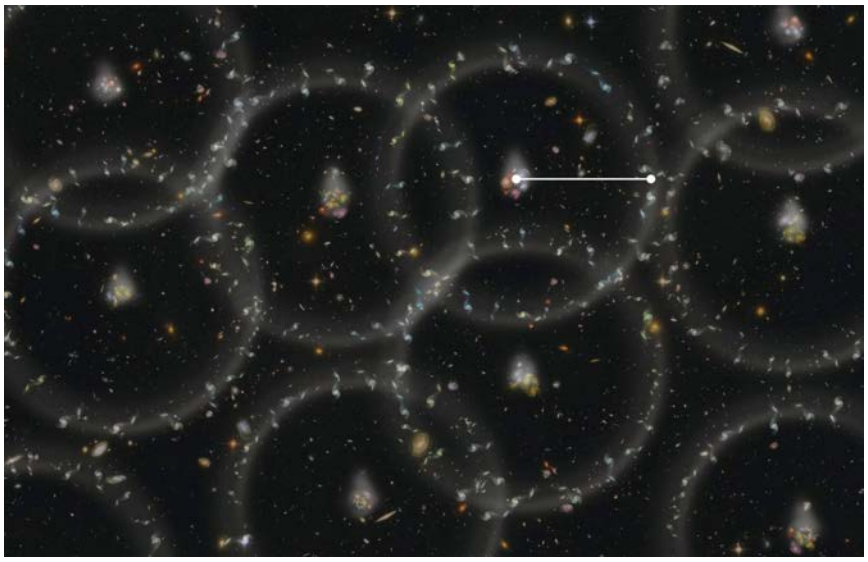

FIGURE 2.5: Cartoon displaying the concept of BAO: around large overdensities, galaxies have a larger probability to be found at BAO distance, since the original sound waves imprinted larger baryonic densities at these scales. The white ruler represents the sound horizon. In this figure, the effect is exaggerated and not to scale. Credit: Zosia Rostomian, Lawrence Berkeley National Laboratory.

ionised (HII), thus being completely transparent to radiation. However, even in the local Universe a fraction of this hydrogen is still in neutral form (HI), which means that when a photon collides, it is absorbed to excite the HI electron. This renders HI effectively opaque to the wavelengths that can excite the electron from its ground state ($n = 1$, with n being the principal quantum number of HI orbitals), to any state with higher energy ($n > 1$).

These wavelengths are known as the *Lyman series*, named after physicist Theodore Lyman. Table 2.1 shows the wavelengths and names of all of its lines, given that the wavelength step between lines decreases at higher n; for $n > 11$ the subsequent lines are virtually indistinguishable from an absorption continuum, referred to as the Lyα limit. All of these lines lie in the UV spectrum, and thus, are blocked by the atmosphere and impossible to observe from ground telescopes unless its source is at redshift high enough ($z > 2.75$).

Even if only a small fraction of the IGM gas is HI, its cross-section is wide enough, and the size of the gas-occupied regions large enough, that there is significant absorption of any light passing through at the Lyα wavelength or higher. Therefore, when observing the spectra of objects in the right wavelengths, these IGM absorption lines should be seen.

Nevertheless, as the radiation emitted by the observed source at z_{source} travels towards us, it will progressively be redshifted due to the expansion of space (see section 1.2.6). Thus, any observer between the source and us will also observe this radiation at a certain z, smaller than the redshift observed by us ($z < z_{\text{source}}$), given that, as an intermediate observer, it is closer to the source.

This has a fundamental consequence: all the spectra originally emitted at bluer wavelengths than the Lyα line in rest frame will, at some point between the source and us, be at the Lyα wavelength, hence becoming susceptible of absorption by the IGM. As a result, the observed spectrum will have multiple emission lines at bluer

TABLE 2.1: Wavelengths and names of the Lyman series.

Name	Lyα	Lyβ	Lyγ	Lyδ	Lyϵ	Lyζ	Lyη	Lyθ	Lyι	Lyκ	Lyman limit
n	2	3	4	5	6	7	8	9	10	11	> 11
λ (nm)	121.6	102.6	97.3	95.0	93.8	93.1	92.6	92.3	92.1	91.9	< 91.2

wavelengths than the Lyα emission at the redshift of the source ($\lambda_{\mathrm{Ly}\alpha} \cdot (1 + z_{\mathrm{source}})$). In fact, each one of these absorption lines, if observed at a wavelength $\lambda_{\mathrm{absorption}}$, will come from an IGM cloud at redshift

$$z_{\mathrm{absorption}} = \frac{\lambda_{\mathrm{absorption}}}{\lambda_{\mathrm{Ly}\alpha}} - 1. \tag{2.35}$$

This ensemble of absorption lines due to the IGM is known as the *Lyα forest*. Generally, it can be observed in the spectrum of quasars at high enough redshift ($z > 2.75$), due to both the extreme luminosity of the quasars (which allows to detect them even at $z > 6$), and the uniformity of its spectrum, even in the UV.

Figure 2.6 shows an example of quasar spectrum with a Lyα forest, together with a cartoon representation of the quasar, the line of sight and the space across the quasar and Earth. All the forest absorption lines correspond to IGM clouds between the quasar and the observer at lower redshifts; the most pronounced absorption line corresponds to the gas surrounding a galaxy that almost intersects the line of sight.

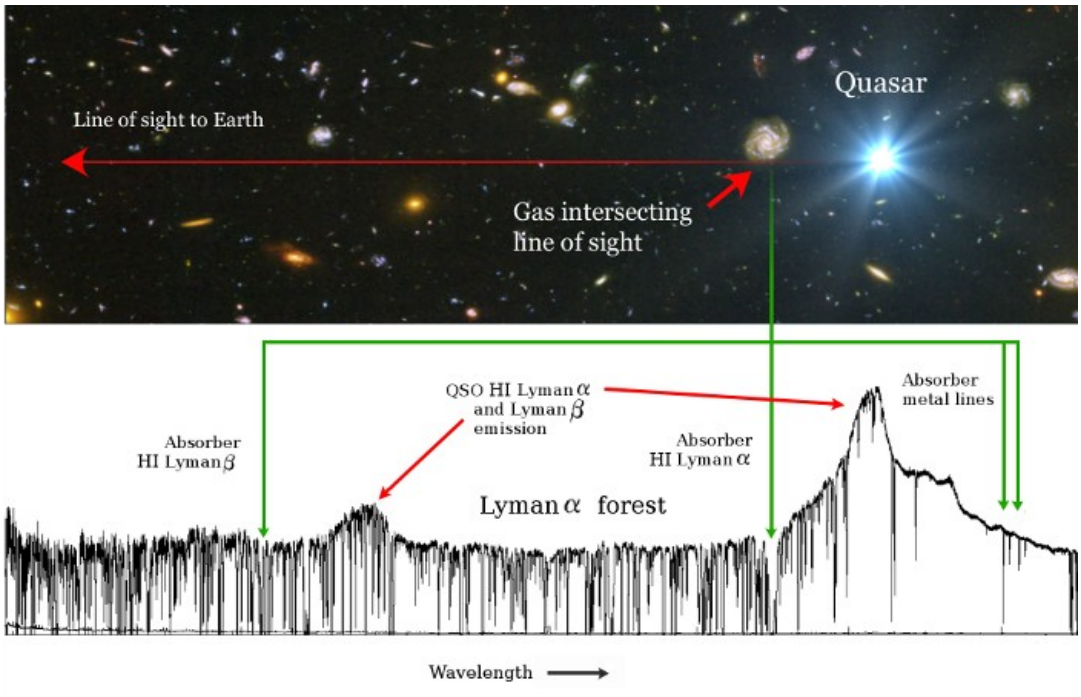

FIGURE 2.6: Cartoon of the Lyα forest. Top panel shows the quasar and its line of sight to earth, with an intersecting gas cloud, in real space. Bottom panel shows the spectra of the quasar in observed frame, with the Lyα line centred in the quasar. Absorption lines to the left (bluewards) of the Lyα line correspond to the Lyα forest. Credit: Neil Crighton.

The Lyα forest was initially hypothesised and detected in Gunn and Peterson, 1965, referred to as the "Gunn-Peterson effect". In the original description this phenomenon was expected to be a uniform absorption of the observed spectrum at shorter wavelengths than Lyα, caused by a rather homogeneous distribution of HI in the IGM. Nevertheless, later observations with higher spectroscopic resolution (Lynds, 1971) showed that the partial absorption was actually a jagged spectrum composed of hundreds of successive absorption lines; hence the "forest" denomination.

2.4.2 Basic properties of the forest

Although the Lyα forest traces the complex underlying dark matter distribution, some of its properties can be approximated by remarkably simple observational laws, while other are heavily dependant on the physical modeling, and present degeneracies that can not be broken by direct observation (e.g., ionisation fraction and temperature). In this subsection we will briefly discuss the three basic parameters that characterise the absorption lines of the forest: its column density, line profile and redshift distribution.

For a thorough review of the subject, we refer the reader to Rauch, 1998, however, we must warn that some of the proposed models for the formation of Lyα absorbers explained there are outdated, and the CDM hierarchical model is accepted as the current paradigm. It is also worth noting that all the features described here are reproduced fairly well (at least with $\sim$10% accuracy) by hydrodynamic simulations;

Column density

In spectroscopy, the *column density* of a certain absorption line is the density of absorbing particles that generates a certain absorption line, in units of density per surface area (usually given in cm^{-2}). Since this refers only to the density of particles absorbing light, whenever column density is discussed for the Lyα forest, it refers solely to HI.

So as to determine the column density of a certain observer, first we need to define the optical depth as a function of the observed flux intensity,

$$\frac{I_{\mathrm{obs}}}{I_{\mathrm{cont}}} = e^{-\tau}. \tag{2.36}$$

Here, I_{obs} is the flux intensity integrated along the absorption line, I_{cont} the flux intensity of the fitted continuum integrated in the same range (i.e., the flux if there was no absorption), and τ the optical depth, a dimensionless parameter which is related to the column density with the following expression

$$\tau = N \cdot \sigma_{\mathrm{a}}, \tag{2.37}$$

where N is the column density in cm^{-2}, and σ_{a} the effective cross-section for the absorption line, in cm^2.

The Lyα forest starts being detectable at column densities of $N \sim 10^{12}$ cm^{-2}, a density low enough to be approximately of the same order of the mean baryonic density of the Universe. The lines become deeper until $N \sim 10^{17}$ cm^{-2}, where the Lyα absorption saturates, but even then absorption systems can be detected up to $N \sim 10^{22}$ cm^{-2} (see section 2.4.3).

Despite such extreme range of column densities (spanning about 10 orders of magnitude), the column density distribution presents a quite straightforward behaviour, that can be modelled at first order as a power law

$$\frac{d\mathcal{N}}{dN} \propto N^{-\beta}, \tag{2.38}$$

where $\mathcal{N}$ is the number of absorbers (absorption lines), and β the exponent of the power law, with values generally around 1.5. Thus, this expression gives the number of absorbers for a given column density; logically, the higher the column density, the least absorbers. It is worth noting that this is just an approximation, and other expressions (e.g., a broken power law with a steeper form for higher densities) may be a better match to the observations.

Line profiles

Given that the IGM clouds of the Lyα forest are relatively simple physical systems (compared to galaxies), its absorption lines closely follow Voigt profiles. Such profiles are the convolution of a Lorentzian and a Gaussian distribution. The first accounts for the homogeneous broadening (caused by the intrinsic uncertainty of the energy levels in the Lyα absorption), while the second represents other physical processes (namely thermal motion or turbulence).

This profile can be can be simply described by two parameters, its peak value (which yields the redshift of the absorber), and the Doppler parameter, $b = \sqrt{2}\sigma$. The latter can be described by a truncated Gaussian distribution of mean $\sim$30 km/s, standard deviation $\sim$15 km/s, and a lower cutoff at $\sim$10 km/s, without a clear correlation with column densities.

The most important (and direct) scientific result that can be obtained from the Doppler parameter is a upper limit on the temperature of the absorber, considering that b follows the expression

$$b = \sqrt{\frac{2kT}{m} + b_{\text{other}}^2}, \tag{2.39}$$

where k is the Boltzmann constant, T the HI temperature, m the mass HI mass, and b_{other} the dispersion caused by other physical effects (mostly turbulence). These upper limits usually yield temperatures of $T \sim 10^4$ K, which is similar to the expected temperature of photoionisation equilibrium, $T > 3 \cdot 10^4$ K. Although photoionisation equilibrium is usually assumed in physical models, the actual temperatures may actually be lower, and different mechanisms for cooling of the HI have been explored.

Distribution of absorbers

If a certain column density threshold is set (be it the observational limit of 10^{12} cm^{-12} or a higher one), the number of absorber per redshift unit can be approximately described by the following power law,

$$\frac{d\mathcal{N}}{dz} \propto (1+z)^{\gamma}. \tag{2.40}$$

This expression, however, shows a clear evolution: as we go further in redshift, the γ exponent also increases. For the relatively low-redshift Lyα forest ($1 < z < 3$, from

the UV to the start of the optical range), generally we find $2 < \gamma < 3$, while at $z \sim 4$ the power law becomes noticeably steeper, with higher γ values.

This increase on the number of absorbers with redshift (and the fact that, for $z \sim 4$ it becomes more drastic) is a confluence of several factors: one one hand, those related to the evolution of the dark matter structure, and on the other hand, the physical processes of collapse and reionization that the absorbers might undergo. Nevertheless, the dominant effect, specially at high redshifts, is the reionization of the IGM: although cosmic reionization is considered complete at $z \sim 6$, there is still a progressive reduction of the residual HI with time.

In addition to its number density with redshift, the distribution of absorbers can also be studied with other approaches. One of them is simply clustering (i.e., the computation of the 2PCF); in that regard, low-column density absorbers ($N < 10^{15}$ cm^{-2}) are very weakly clustered (much less than galaxies), while for absorber of higher column densities the 2PCF increases. This may be a hint of the denser absorbers being galaxy precursors, or at least being placed in gravitational wells from dark matter halos. More results from the clustering measurements of the Lyα forest will be discussed in section 2.4.4.

Another feature worth mentioning of the absorber distribution is the *proximity effect*: the reduction on the number of absorbers as we approximate to the redshift of the quasar. While this adds an extra layer of complexity when modeling the forest (e.g., can other quasars close to the line of sight artificially decrease the absorber density?), it can also be used to estimate the UV background in the IGM (Carswell et al., 1987).

2.4.3 Special features

The characteristics explained in the previous subsection describe fairly well the observed Lyα forest for most quasars ($z < 6$), and absorbers of low enough column density ($N < 10^{17}$ cm^{-2}), as long as no other relevant IGM lines appear in the same wavelength range as the forest. Here, we will briefly discuss the elements that appear in the forest spectrum when one of these conditions is not met.

Lyman limit systems

In principle, the relation between optical depth and column density is linear (eq. (2.37)), and an increase in N only results in an increase of the absorbed flux, without altering the shape of the absorption line in the spectrum.

These assumptions, however, do not hold true anymore when absorption is strong enough to make the absorption line reach zero flux at its peak; in such case, the absorber becomes effectively opaque to the Lyα emission from its reference frame. An increase of N will modify the shape of the line, expanding the region where the continuum is totally absorbed, and making the transition between the continuum and the opaque absorber more abrupt, close to a discrete step. This is referred to as the saturation of the absorption line.

In the Lyα forest, an absorber saturates completely when $N > 10^{17}$ cm^{-2}. These absorbers are called *Lyman limit systems*, due to the fact that they become also opaque in the Lyα continuum. Figure 2.7 displays an example of a Lyman limit system from one of the earliest studies of these absorbers (Prochaska, 1999).

Due to the saturation, in these systems the column does not show an straightforward relationship as in eq. (2.37). The total absorbed flux becomes dependant on $\log N$, and a degeneracy appears with the Doppler parameter, b (given that the shape

of the absorption line can not be constrained anymore regardless of its area). Nevertheless, given the resolution of modern spectrographs, N can still be determined by profile fitting.

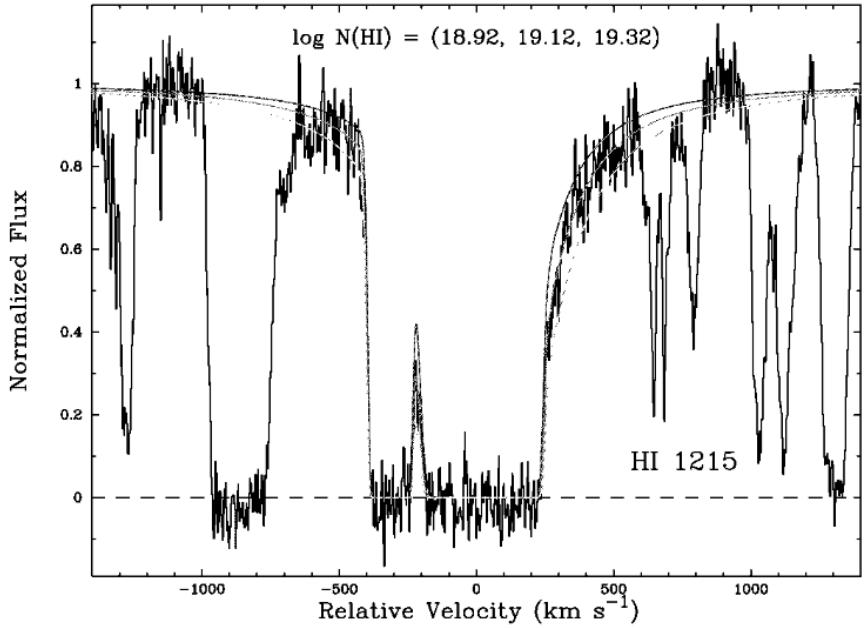

FIGURE 2.7: Example of Lyman limit system, detected at $z = 2.652$, but with the x scale centred in the system for this figure. Smooth lines represent fits for different column densities of the system. From Prochaska, 1999, Fig. 1.

Damped Lyα absorbers

If the column density of an absorber increases even more ($N > 2 \cdot 10^{20}$ cm^{-2}), the saturated absorption line broadens, but its profile becomes again less abrupt. The continuum in the vicinity of the absorption line starts decreasing in a smooth slope that becomes steeper close to saturation, but without the vertical jump of the Lyman limit systems. These features are referred to as *damping wings*, and thus the absorbers that exhibit this behaviour are *Damped Lyα absorbers*. Figure 2.8 shows an observational example.

The origin of these damping wings lies in the intrinsic uncertainty of the lifetime of excited HI, which in turn results in a uncertainty in its energy level, and thus a broadening of the absorption line. For unsaturated Lyα absorbers this effect is combined with the Doppler effect from the motion of the gas, resulting in the aforementioned Voigt profile.

Nevertheless, deviations too large from the average wavelength of the line are only possible due to this lifetime broadening; on such regions, this broadening will dominate over the Doppler effect. However, the likelihood of such deviations is small enough, so in unsaturated absorbers, and even Lyman limit systems these regions are optically thin. Only when N increases well above saturation this broadening due solely to lifetime uncertainty becomes apparent, resulting in the appearance of damping wings.

In these damped systems, the equivalent width of the absorption is proportional to $\sqrt{N}$, without degeneracies with b. This makes damped Lyα absorbers more informative and easier to study than Lyman limit systems, especially with the lower resolution spectrographs used in early works on the subject (e.g., Carswell et al., 1975).

In addition to this, damped Lyα absorbers show some physical properties that make them the most studied feature of the Lyα forest. Given its high column density, inside an outer shell of ionised gas most of the hydrogen in damped Lyα absorbers remains as HI, making them the largest reservoirs of neutral gas in the Universe at $z < 5$. This is of special importance, since HI is an essential element for star formation.

Besides, since a larger gravitational potential is needed to maintain these systems bound, they are believed to reside in small halos, and show more clustering than standard Lyα forest lines. Therefore, damped Lyα absorbers play a crucial role in galaxy formation and evolution, and have been largely studied in these fields. For a comprehensive review on that matter, the reader may refer to Wolfe, Gawiser, and Prochaska, 2005.

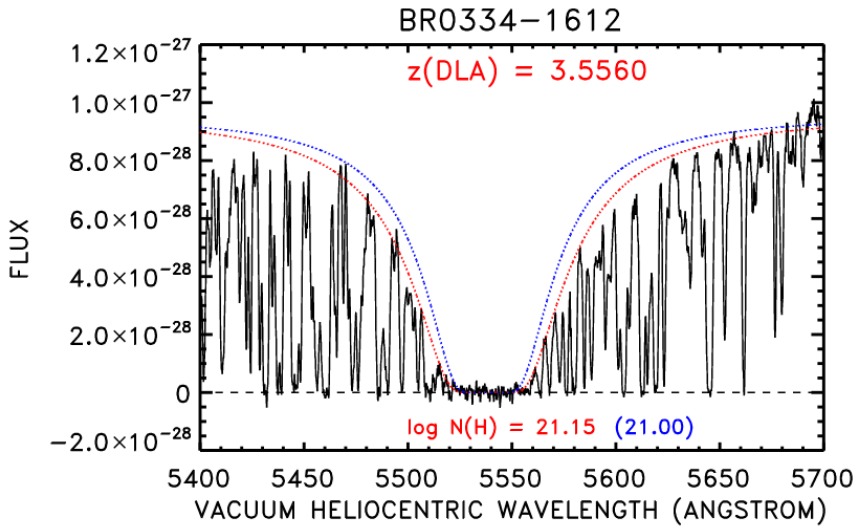

FIGURE 2.8: Example of damped Lyα at $z = 3.556$; colored lines represent fits for different column densities. From Songaila and Cowie, 2010, Fig. 7.

Gunn-Peterson trough

So far, we have seen the elements that appear in a quasar spectrum when a certain region of the IGM is dense enough to become effectively opaque to Lyα emission. If the HI density was high enough for saturation in all of the IGM, everything bluewards from the Lyα line of the quasar would be totally absorbed.

This is exactly what should happen before reionisation, and is referred to as *Gunn-Peterson trough*. Hypothesised in Gunn and Peterson, 1965, it was finally discovered in Becker et al., 2001 analysing the spectrum of a SDSS quasar at $z = 6.28$ (fig. 2.9). This was one of the first direct observational probes of reionisation, and was critical in determining its end at $z \sim 6$.

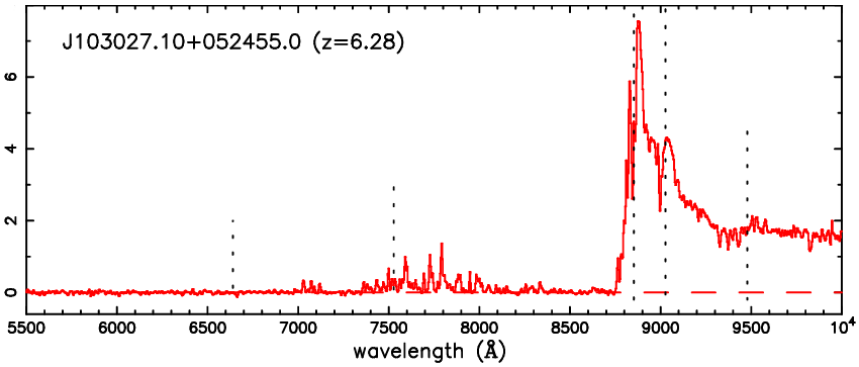

FIGURE 2.9: One of the first detections of the Gunn-Peterson trough, at $z = 6.28$. Going in descendent wavelength direction, the trough begins almost immediately after the Lyα line. From Becker et al., 2001, Fig. 1.

Lyβ forest

The HI in the IGM can absorb at any Lyman wavelength from table 2.1, although Lyα is the most common transition, given that it is the least energetic. Therefore, in the spectrum of a quasar a forest of absorption lines will appear bluewards from the Lyβ line, the next one in the series.

Nevertheless, the Lyβ line is already will already be inside the Lyα forest region of the spectrum (since it is bluer than Lyα), so the result will be a combination of both the Lyα and Lyβ forests after the Lyβ line, each one coming from a different redshift. For a given observed wavelength $\lambda_{\rm obs}$, the Lyα forest absorption will come from a redshift $z_{Ly\alpha} = \lambda_{\rm obs}/\lambda_{Ly\alpha} - 1$, while the Lyβ forest absorption will happen at $z_{Ly\alpha} = \lambda_{\rm obs}/\lambda_{Ly\alpha} - 1$.

This superimposition of two signals with a large redshift offset greatly limits the usefulness of the Lyβ forest, as the Lyα and Lyβ absorbers are far away enough to have no correlation (i.e., far above the scale of homogeneity). In addition, the redshift range where the Lyα forest can be fully exploited is also limited to the region of the spectrum between the Lyα and Lyβ lines. However, the Lyβ forest region can still be used for cosmological studies, mainly because Lyα forest cross-correlation can still yield results in spite of the Lyβ signal (e.g., Blomqvist et al., 2019).

2.4.4 Cosmology with the Lyα forest

The Lyα forest is a valuable cosmological tool, as it traces the underlying dark matter structure. Nevertheless, the relationship between the observed optical depth τ in the forest absorbers and the overdensity δ is far from trivial, and can not be separated from other parameters, both cosmological and relative to physical conditions of the IGM. This is illustrated by the Fluctuating Gunn-Peterson Approximation (FGPA, Weinberg et al., 1997), which can be expressed as

$$\tau_{\rm HI}(z) = 0.172 h^{-1}(1+\delta)^{2-0.7\gamma}\left(1+\frac{1}{H(z)}\frac{dV_{\rm los}}{dx}\right)^{-1}\times \left(\frac{1+z}{4}\right)^{6}\left(\frac{H(z)}{5.51H_0}\right)^{-1}\left(\frac{\Omega_b h^2}{0.0125}\right)^{2}\left(\frac{T_0}{10^4}\right)^{-0.7}\left(\frac{\Gamma}{10^{-12}}\right)^{-1}. \tag{2.41}$$

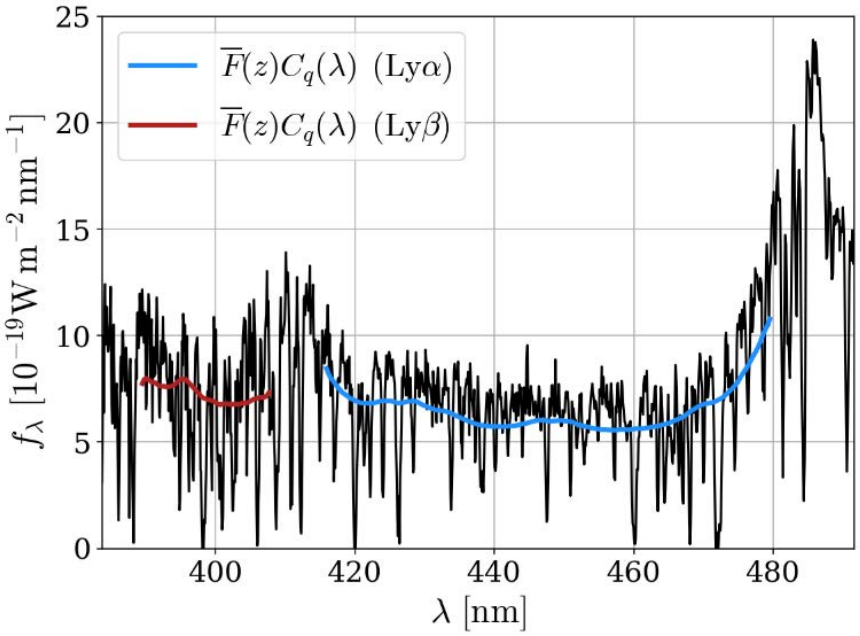

FIGURE 2.10: Detail of the Lyα and Lyβ forest in a quasar spectrum at $z = 3$. Blue and red lines represent the continuum fit for the Lyα and Lyβ forest, respectively; the bump in the spectrum between both (around ∼410 nm) is the Lyβ line emitted by the quasar. From Blomqvist et al., 2019, Fig. 3.

In addition to several cosmological parameters already covered in 1.2, other variables of astrophysical origin appear. T_0 (in K) and γ are the parameters of a power law that relates HI temperature to overdensity

$$T(\delta) = T_0(1+\delta)^\gamma, \tag{2.42}$$

while Γ is the photoionisation rate, in s^{-1}. Besides, the dV_{los}/dx term is the spatial derivative of the line-of-sight velocity, which accounts for the differences between real space and redshift space (since the physical modelling is performed in the former, but the latter is observed).

This FGPA assumes that eq. (2.42) holds, and that gas density is low enough for pressure gradients to become negligible and gravity dominate, which makes HI overdensity be close to the actual dark matter δ. Therefore, it breaks $\delta \gtrapprox 10$, which does not include most of the forest lines ($N \gtrapprox 10^{16}$). In this regime, if all the parameters but δ are assumed known, the relation between τ and δ becomes a simple power law, but the physical properties of the Lyα forest are far from certain.

In spite of these degeneracies with IGM properties, the Lyα forest as a cosmic tracer still presents some distinct advantages. First, it is one of the best clustering probes in its redshift range ($2 < z < 4$), since at these high redshifts galaxy density in surveys is much smaller than at $z < 2$ (given that only the brightest are observed). Second, it traces really well the low overdensity regions at relatively large scales, which are mostly in the linear regime. These features makes the Lyα forest suitable to constrain warm dark matter models (WDM) or neutrino masses, which both should cause suppression and cut-offs in the power spectrum (Viel, 2003). Therefore, since its first observations the Lyα forest has been used as a tool to both measure clustering and place constraints on the aforementioned physical models.

Some recent examples of clustering measurements are the determination of the 1D power spectrum (along the line of sight) in Chabanier et al., 2019, as well as the 3D power spectrum and BAO in Bautista et al., 2017, both with eBOSS data releases. While the 1D power spectrum can be determined with relatively few quasars, its three-dimensional counterpart requires a large enough quasar density, since with the scarce space sampling of the Lyα forest the power spectrum of the whole volume is being determined. Regarding the model constraints, for example Yèche et al., 2017 places constraints in neutrino masses using the 1D power spectrum from the Lyα forest, and Garzilli, Boyarsky, and Ruchayskiy, 2017 discusses the degeneracies between warm and cold dark matter models and thermal histories of the IGM when analyzing Lyα forest spectra.

Chapter 3

Cosmological surveys and Intensity Mapping

The task of systematically registering the positions of celestial bodies has been carried out by most civilisations through History, with the Hipparchus star catalogue (compiled in ∼135 BC) being the most famous historical example. However, the conception of extragalactic bodies, such as other galaxies than the Milky Way, dates from the beginning of the XXth century (see section 1.1.3), and the observation of all but the closest of these galaxies requires modern telescopes. Therefore, the elaboration of catalogues of galaxies with cosmological purposes is a recent scientific activity, with the first of them being the Center for Astrophysics Redshift Survey (Davis, Huchra, and Latham, 1983). These kind of observations, as well as its derived catalogues, are referred to as *cosmological surveys*.

In the broadest sense, cosmological surveys are observational programs designed to systematically scan certain areas of the sky, with the aim of obtaining an uniform mapping of a significant fraction of the Universe at certain wavelengths and redshift. While surveys can be carried out in any wavelength of the electromagnetic spectrum, the surveys considered in this thesis are only in the optical part of the spectrum (and thus, their observations are performed with conventional ground telescopes).

Optical surveys can be classified in two large groups: imaging surveys and spectroscopic surveys. In the first, sky images are directly registered with a CCD camera mounted on the telescope, and a set of filters in front of the camera determine the wavelength ranges that are observed, given by the response function of each filter (effective throughput of the filter versus wavelength, e.g., see fig. 3.3). In the second case, an spectrograph is mounted on the telescope, which disperses the observed light depending on its wavelength by diffraction. This dispersed light is then registered by different pixels of a CCD, thus obtaining a high-resolution continuum of observed flux versus wavelength (e.g., fig. 2.9), instead of a net flux value per each filter.

While spectroscopic surveys provide far more detailed information of the flux emitted by an object, imaging surveys have the advantages of not requiring previous object selection (as the whole sky is observed, instead of specific targets), and requiring much less observation time to achieve higher SNR and observe fainter objects (since each pixel of the CCD receives photons from a much larger wavelength range).

In this chapter, we will first discuss the observational technique that will be evaluated for the the Lyα line in this thesis: Intensity Mapping (IM). Then, we will delve into the characteristics of the surveys simulated in this work: a narrow-band imaging survey, the Physics of the Accelerating Universe Survey (PAUS), the Extended Baryon Oscillation Spectroscopic Survey (eBOSS), and the Dark Energy Spectroscopic Instrument (DESI).

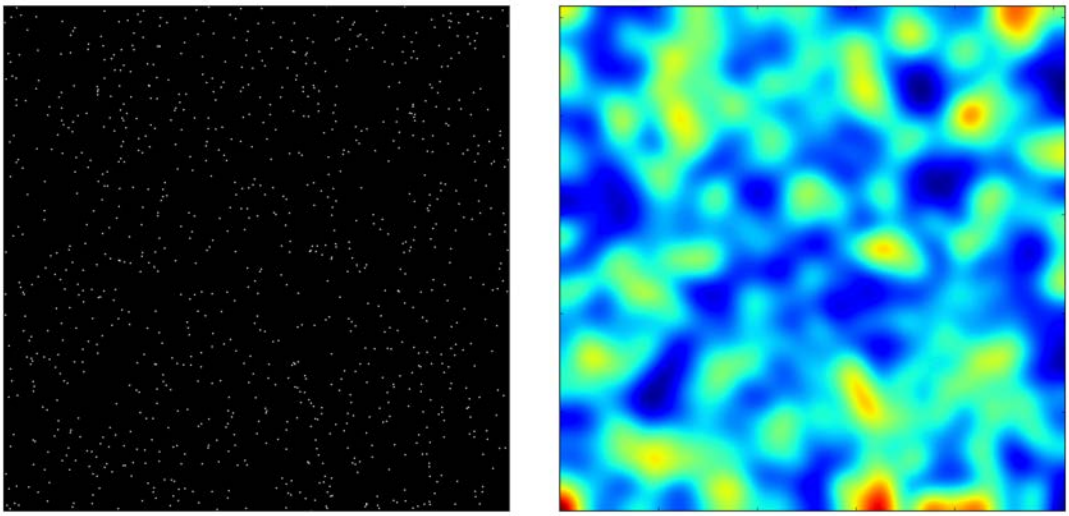

FIGURE 3.1: *Left panel:* galaxy positions in a simulation. *Right panel:* HI distribution corresponding to the simulated galaxies. Credit: Francisco Villaescusa-Navarro.

3.1 Intensity Mapping

In the last few years, the amount of observational data for the Universe at different wavelengths has steadily increased, which has led to the development of new methods and techniques to analyse these observations. IM is one of these techniques, consisting of the tracing of large-scale structure with one or more emission lines, without resolving any kind of finite source, like galaxies or quasars. The use of a sharp and narrow spectral feature, such as an emission line, allows us to map the structure not only in angular coordinates but also in redshift, which provides a 3D tomography of the tracer (Peterson et al., 2009).

Figure 3.1 shows the difference between IM and conventional observation of resolved objects: on the left panel there are displayed as dots the resolved galaxies from a simulation, while in the right panel the HI distribution corresponding to these galaxies is shown. This is the same HI in the IGM originating the Lyα forest (section 2.4), and as this HI radiates with different emission lines (e.g., 21 cm, Lyα), with IM the distribution of the right panel would be observed, instead of the left one.

Originally, this technique was proposed to study the power spectrum with the 21-cm emission line at high, pre-reionization redshifts ($z > 5$) (Madau et al., 1997; Loeb and Zaldarriaga, 2004), but its application at lower redshifts has also been studied, e.g., as a method to measure BAO (Chang et al., 2008). Other emission lines have also been considered, such as the CO rotational line at intermediate (Breysse, Kovetz, and Kamionkowski, 2014; Li et al., 2016) or high redshift, (Carilli, 2011), CII emission line (Gong et al., 2012; Yue et al., 2015), or the Lyα line (Silva et al., 2013; Pullen, Doré, and Bock, 2014). Given the relatively short wavelength of Lyα emission, this line can only be observed at $z > 2$ with ground-based telescopes, which limits any IM study with this tracer to relatively high redshifts.

Since IM does not resolve individual objects but considers all emission at certain wavelengths, one of the main challenges that IM studies face is contamination by foregrounds. It is worth noting that, for most cases (i.e., all lines but Lyα, and

other tracers that may have been used for cross-correlation), IM is a technique applied to radio telescope observations, as these emission lines go from the far infrared well intro the radio spectrum. Therefore, both the observational data and the analysis techniques are completely different: radio telescopes have much less angular resolution but much larger angular coverage, and the treatment of both foreground contamination and other noise sources is fundamentally distinct.

Regarding foregrounds, this source of noise can be removed via cross-correlation with other datasets of objects with well-known redshift, an approach that has been successfully applied in detections of the 21-cm line (Chang et al., 2010), CII emission line (Pullen et al., 2018) and the Lyα line (Croft et al., 2016; Chiang, Ménard, and Schiminovich, 2019). We will discuss the particularities of the few Lyα IM studies, since they are the closest to our case (both because they deal with the same emission line, and the bulk of their data comes from optical/UV telescopes, not radio ones).

In the Lyα case, in Chiang, Ménard, and Schiminovich, 2019 detection of extragalactic background light (EBL) is reported using cross-correlation of UV broadband data from the Galaxy Evolution Explorer (GALEX, Gil de Paz et al., 2007) with spectroscopic galaxy samples; with an adjustable spectral model of the EBL they place constraints on total Lyα emission up to $z = 1$. This work only considers the evolution of EBL and its properties in redshift direction by integrating the cross-correlation in an angular range corresponding to 0.5-5 Mpc/h, so other than the redshift evolution, the results of this work are confined to cluster scales.

On the other hand, in Croft et al., 2016 all data used for IM was extracted from the Sloan Digital Sky Survey III (SDSS-III, Eisenstein et al., 2011) BOSS (section 3.3). Lyα emission is estimated by selecting spectra of Luminous Red Galaxies (LRGs) at $z < 0.8$ and subtracting a best fit model for each galaxy spectrum, which leaves a significant amount of Lyα surface brightness from higher redshifts. These residual spectra are cross-correlated with quasars from the same catalogue, which gives a detection at mean redshift $z = 2.5$ of large-scale structure at a 8σ level, and a shape consistent with the ΛCDM model. This cross correlation, however, only yields a positive signal on scales up to 15 Mpc/h. Given the quasar density of BOSS, this implies that only 3% of the space (15 Mpc/h around quasars) is being mapped, and large scale structure of Lyα emission in general is not being constrained by this measurement. Lyα emission is extended at high enough redshift (approximately $z > 3$), with Lyα blobs (Taniguchi, Shioya, and Kakazu, 2001; Matsuda et al., 2004) forming visible structures around quasars up to hundreds of kpc in size, and the integrated faint Lyα emission in turn covers almost 100% of the sky (Wisotzki et al., 2018). Therefore, cross-correlation of the Lyα emission with a more suitable dataset (less rare than quasars) is expected to provide a positive signal on larger scales.

One of these possible datasets to cross-correlate with is the Lyα forest, as the Lyα emission observed with IM comes from the very same IGM that causes the Lyα forest. As described in section 2.4, each Lyα forest spectrum contains information about the HI distribution along a large fraction of the entire line of sight, which should allow cross-correlation over larger, more representative volumes. In Croft et al., 2018 a first attempt at cross-correlation was performed between Lyα forest from BOSS and similar LRG spectra with the best galaxy fit subtracted to those used in Croft et al., 2016, but no signal was found. Nonetheless, BOSS was not designed with Lyα IM as an objective, and it is certain that larger and more suitable datasets are needed to obtain a clear detection (Kovetz et al., 2017). Such a dataset would need data with redshift precision close to that achieved by spectroscopy over large areas, providing a volume large enough to study large-scale structure with Lyα IM (Croft et al., 2018).

One potential candidate that may fulfil these requirements are narrow-band imaging surveys, such as PAUS. The objective of this work is therefore to simulate a study of IM with the Lyα line using the images from PAUS, whose background (which contains Lyα emission from the IGM) is to be cross-correlated with the Lyα forest date from eBOSS and DESI. The characteristics of these surveys will be briefly described in the following sections.

3.2 PAUS

Most imaging surveys observe the night sky using broad-band filters: these filters have wide response functions, usually with FWHM>100 nm, so with a few of them the whole optical spectrum is covered. With such filters, the spectral information of the observed objects is fairly limited (since we only get a few points of the optical spectrum), but good SNR can be obtained with relatively low observation times. Some examples of these surveys are the Sloan Digital Sky Survey (SDSS, Ahumada et al., 2020) or the Dark Energy Survey (DES, Abbott et al., 2018).

One of the main purposes of these surveys is the determination of the redshifts of the observed objects knowing their fluxes in each filter (usually known as *photometric redshifts*). In spite of the very limited spectral information provided by the broad-band filters, our current knowledge of galaxy spectra, and the calibration with complementary spectroscopic observations, allows to determine photometric redshifts with errors of the order of few percent for objects at $z < 1$ (e.g., Hoyle et al., 2018).

Nevertheless, narrow-band surveys, with filters significantly narrower than the standard broad-band ones, have also been carried out. They offer a trade-off between broad-band and spectroscopic surveys: higher spectral resolution than broad-band, but also higher observation times. The Physics of the Accelerating Universe Survey (PAUS), in which this work is focused on, is an state-of-the-art narrow-band imaging survey, currently being carried out at the William Herschel Telescope in Roque de los Muchachos, with the PAU Camera (Castander et al., 2012). Figure 3.2 displays the instrument already mounted and ready for observation, as well as the William Herschel Telescope.

The main feature of this instrument is the use of 40 narrow-band filters with a full width at half maximum (FWHM) of $\simeq$ 13 nm, with mean wavelengths of 455 to 845 nm in steps of 10 nm. Figure 3.3 displays the response function of the 40 PAUS filters (colored), together with the standard *ugriz* bands from SDSS (black lines). Such configuration allows one to obtain photometric redshifts (photo-z) with sub-percent precision over large sky areas (Martí et al., 2014). Preliminary results (Eriksen et al., 2019) already achieve better photo-z precision than state-of-the-art photo-z measurements in the COSMOS field (Laigle et al., 2016).

The main purpose of the survey is the elaboration of high-density galaxy catalogues with high-precision redshifts for cross-correlations of lensing and redshift distortion probes (Gaztañaga et al., 2012). As of the time of writing, the latest public catalogue (Alarcon et al., 2020) is complete in the COSMOS field up to $i_{\rm AB} < 23$, and at the completion of the survey similar catalogues are expected to be released in the other fields observed by PAUS. These are the W1, W2 and W3 fields from the Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS, Heymans et al., 2012). Together with COSMOS, they cover approximately $\sim$130 deg^2 of the sky, however, since a full coverage of the CFHTLenS fields is not expected a total angular area

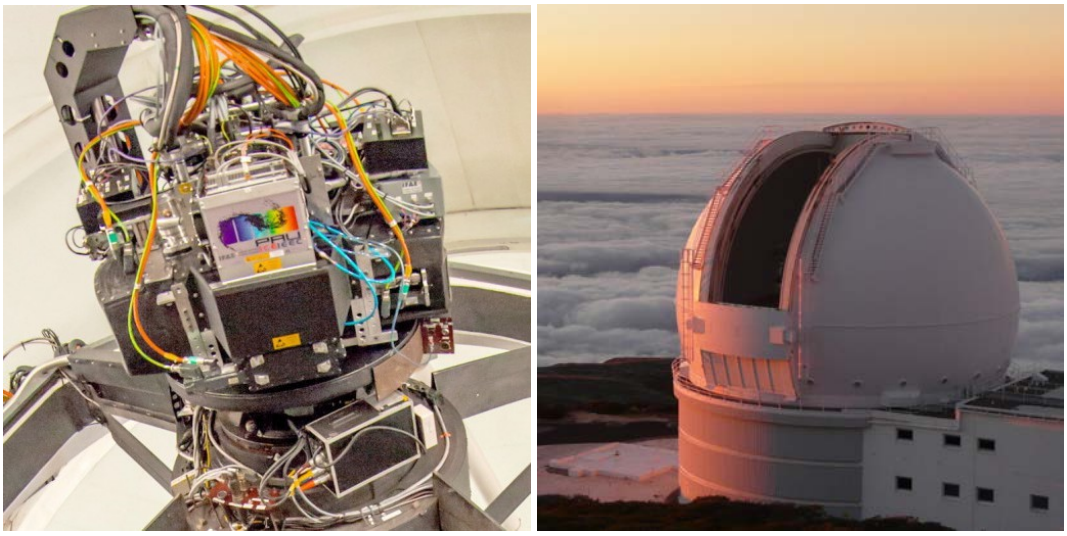

FIGURE 3.2: *Left panel:* PAUS Camera mounted on the William Herschel Telescope. Credit: PAUS Collaboration. *Right panel:* Outer view of the telescope. Credit: 2pem/Wikimedia Commons.

of $\sim$100 deg^2 of PAUS images will be considered for the simulations presented this work.

In addition to these photometric catalogues, the narrow-band data from PAUS may also be used for intensity mapping of Lyα. The background of PAUS images, where no objects are resolved, also contains valuable cosmological information. Given the wavelength range of the NB filters, Lyα luminosity is observed in the range $2.7 < z < 6$, distributed in 40 redshift bins, one per each NB filter. At this redshift range faint Lyα emission surrounds most objects (Wisotzki et al., 2018), but foreground contamination must be removed first in order to study it, since the weak Lyα emission coming from the IGM is several orders of magnitude fainter than the observed galaxies.

In spite of the wavelength coverage of PAUS, however, only the seven bluest NBs will be considered for this work, which span from 455 to 515 nm (shaded in fig. 3.3). With these seven blue filters, Lyα emission is observed over the range $2.7 < z < 3.3$, approximately. At higher filter wavelengths, the observed Lyα emission increases in redshift, thus being farther away and fainter. In addition to this, the fraction of quasars observed at $z > 3.3$ is extremely small (fig. 3.7), which means that the amount of Lyα forest data sampling this space is also very limited. Therefore, adding extra filters only provides a volume for the cross-correlation with lower SNR in PAUS images, and scarcely sampled by the Lyα forest.

3.3 eBOSS

The Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al., 2013) is a survey, part of the larger SDSS project, with the main aim of measuring BAO (see section 2.3) with the spectra of LRGs in a redshift range $z < 0.7$, up to magnitude $i_{\rm AB} < 20$. However, other of the targets of the survey is high-redshift quasars, and their Lyα forest. Given that the Lyα line enters the optical spectrum at $z > 2$, these quasars are observed in the redshift range $2.15 < z < 3.5$, up to magnitudes as faint as

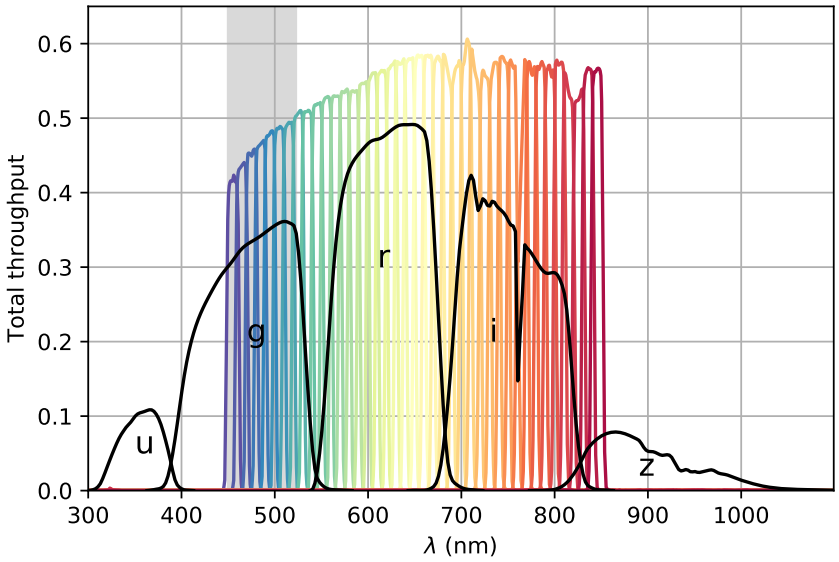

FIGURE 3.3: Response function for PAUS filters (coloured) and original SDSS *ugriz* filters (black). Shaded area represents the wavelength range studied in this work.

$g_{AB} < 22$. The total angular coverage of the survey is of 10,000 deg^2, containing a total of approximately 1.5 million LRGs and 150,000 quasars.

After finishing its observational run, new spectroscopic targets were selected to continue the survey, under the name of the Extended Baryon Oscillation Spectroscopic Survey (eBOSS, Dawson et al., 2016). Most of these new targets come from two populations unexplored by BOSS: 250,000 LRGs at a higher redshift range ($0.6 < z < 2.2$), and 500,000 low-redshift quasars ($0.9 < z < 2.2$). While most of the observed spectra of these quasars does not contain their Lyα forest (as the Lyα line falls well into the UV spectrum), their redshift measurements alone already provide valuable information to study clustering, as they are highly biased tracers. In addition to these two new spectroscopic samples, eBOSS also targets 60,000 new quasars at $z > 2.1$, whose observed spectra indeed contain the Lyα forest. All these new observations are carried out with approximately the same survey footprint, and thus angular coverage; eBOSS increases the total object density in the same fields as the original BOSS catalogue.

The BOSS/eBOSS instrument (Smee et al., 2013) is a multifibre spectrograph that allows the observation of 1,000 targets in a single exposure. In order to do so, each fibre collects the light of a single object, and is rerouted to the spectrograph, where the light emitted by the object is diffracted and observed. These fibres are all concentrated in a fibre cartridge (fig. 3.4, left panel); on top of the cartridge, an aluminium plate drilled at the positions of the selected targets is mounted (fig. 3.4, right panel). Each one of the fibres is manually inserted in the corresponding hole of the plate, thus ensuring that each fibre observes its designated object.

For this instrument, two twin spectrographs are connected to the fibre cartridge, each one with a blue and a red camera, observing different wavelength ranges. The light is diffracted by a grism (a combination of a diffractive grating and a prism), reaching spectral resolutions of $1,560 < R < 2,270$ in the blue channel, and $1,850 < R < 2,650$ in the red channel (with the spectral resolution being defined as $R \equiv \lambda/\Delta\lambda$). The ensemble of fibre cartridge with plate+spectrographs is mounted on the

telescope, with the fibre plate (and the connected fibres) on its focal plane. The telescopes where the observations took place were the Sloan Foundation 2.5m telescope at the Apache Point Observatory for BOSS, and the Irénéé du Pont Telescope (also 2.5m) at Las Campanas Observatory for eBOSS.

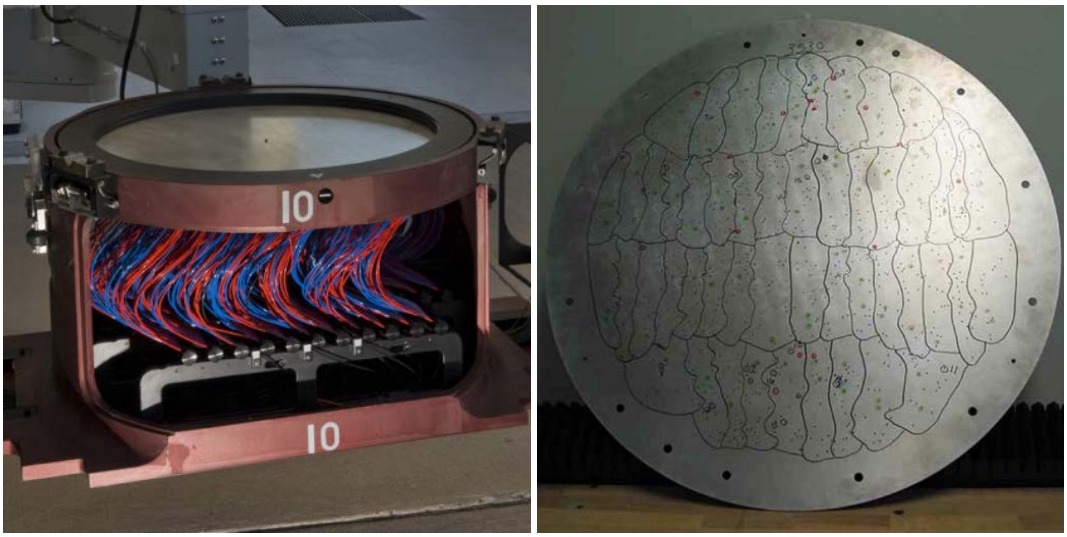

FIGURE 3.4: *Left panel:* Fibre cartridge of the BOSS spectrograph. *Right panel:* Plate for the BOSS spectrograph. Credit for both: SDSS.

Regarding its contribution to science, both BOSS and eBOSS have been the largest spectroscopic surveys at their times, which makes their data invaluable for cosmology. In the particular case of the Lyα forest, most of the recent cosmological results, such as measurements of power spectrum or BAO, have been a product of these surveys (see section 2.4.4).

Given that the spectral resolution of eBOSS is far larger than the narrow-band photometry from PAUS (the maximum resolution achievable for the latter would be around 65 if its photometric data was to be compared against spectroscopy), and that the much larger angular coverage fully overlaps with the PAUS fields, the only parameter we need from eBOSS to model the cross-correlation is the angular quasar density versus redshift of its catalogue. This is necessary to introduce the exact amount of quasars in the simulation, which in turn will determine how much space is actually sampled by their Lyα forests. Figure 3.7 shows this data (orange line for eBOSS), extracted from Dawson et al., 2016, Table 1.

3.4 DESI

The Dark Energy Spectroscopic Instrument (DESI, DESI Collaboration et al., 2016b), is a next-generation spectroscopic survey, with the main goal of constraining different dark energy models by measuring BAO (section 2.3) and redshift space distortions (section 5.3.2). In order to do so, four different samples will be observed: LRGs up to $z < 1.0$, galaxies with bright OII lines up to $z < 1.7$, a bright galaxy sample at $z \sim 0.2$ (to be observed mostly in bright moon phases), and a high-redshift quasar sample at $2.1 < z < 3.5$. In total, more than 30 million objects will be observed.

Regarding its angular coverage of the sky, it will be of 14,000 deg^2, completely overlapping with the eBOSS fields. Besides, fainter targets will be selected, which increases the net object density. In the case of quasars, eBOSS selected targets up to

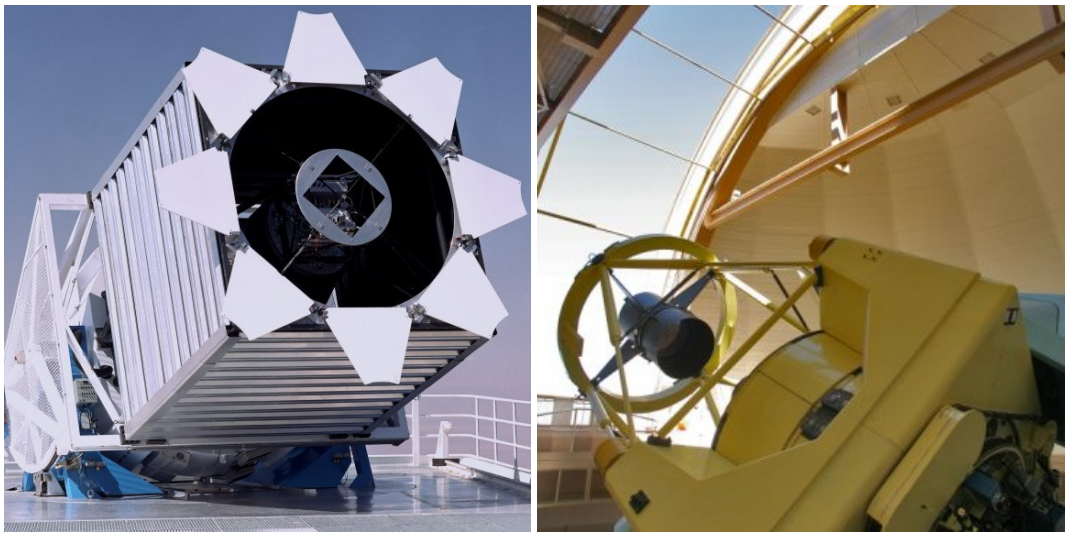

FIGURE 3.5: *Left panel:* Sloan Foundation Telescope at Apache Point Observatory. *Right panel:* Irénéé du Pont Telescope at Las Campanas Observatory. Credit for both: SDSS.

$g_{\rm AB} < 22$, while DESI goes deeper, extending the magnitude limit up to $g_{\rm AB} < 22.5$. Therefore, DESI is considered a direct successor to eBOSS.

The entirety of the survey will be carried out at the Mayall Telescope at the Kitt Peak National Observatory. This is a 4-meter telescope (thus, larger than the 2.5-meter telescopes used in BOSS/eBOSS) that has been completely repurposed for DESI, and will dedicate all of its observing time to the survey. A picture of the telescope, already prepared for DESI, is displayed in fig. 3.6, right panel.

The instrument (DESI Collaboration et al., 2016a) mounted on the telescope will also be a multifibre spectrograph, but instead of using drilled plates to locate the fibres in the focal plane, each fibre will be directed by a robotic positioner, making the process completely automatic. Up to 5,000 targets can be observed in a single exposition (one per fibre), and the observed light is redirected by each fibre to one of the 10 identical spectrographs of the instrument. These spectrographs have three cameras that observe different parts of the spectrum: infrared, red and blue. Their spectral resolution is $2,000 < R < 3,200$ for the blue channel, $3,200 < R < 4,100$ for the red channel, and $4,100 < R < 5,000$ for the infrared channel. Figure 3.6, left panel, displays the focal plane system, showing the robotic positioners with the fibres attached.

As of the time of writing, DESI has finished commissioning, i.e., the tests to verify that the instrument, once mounted, satisfies all the requirements for the survey. The survey validation phase is to be started, and by the next year the acquisition of scientific data should be on its way. More precisely, 1% of the survey is expected to be completed in the first half of 2021.

Regarding the data needed to simulate the cross-correlation in this work, as in the case of eBOSS, the only necessary parameter is the quasar density versus redshift. Its values are displayed in fig. 3.7, blue line, as extracted from DESI Collaboration et al., 2016b, Fig. 3.17.

FIGURE 3.6: *Left panel:* Detail of the DESI focal plane system. Each axis in the focal plane is a robotic positioner with a fibre attached. *Right panel:* Mayall Telescope at Kitt Peak National Observatory. Credit for both: DESI Collaboration.

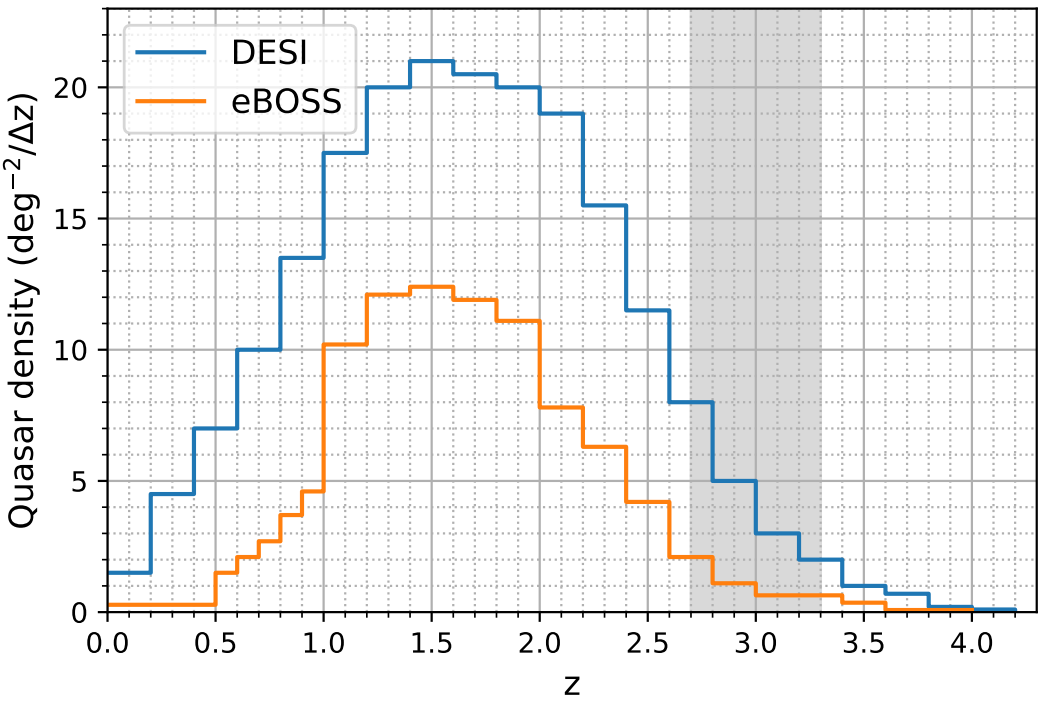

FIGURE 3.7: Projected quasar density vs redshift for DESI and eBOSS; shaded area shows the redshift interval of this study. The change of redshift binning at $z = 0.1$ in eBOSS (from $\Delta z = 0.1$ to $\Delta z = 0.2$) is not relevant for this work.

Chapter 4

Simulation of survey data

As we have stated in section 3.1, the goal of this thesis is to simulate the cross-correlation between PAUS images and the Lyα forest from eBOSS/DESI, with the objective of detecting the extended Lyα emission of the IGM.

In order to simulate this cross-correlation between different surveys, the first step is to simulate the actual survey datasets. The simulation of the background of PAUS images poses a bigger challenge, since independent elements from the Lyα signal need to be taken into account (namely, foreground emission and instrumental/atmospheric noise). On the other hand, the modelling of the observed Lyα forest is quite straightforward, as it already traces the underlying dark matter distribution fairly well, even without any kind of additional data reduction.

For this work, an already existing hydrodynamic simulation has been used for both Lyα forest data and Lyα emission, while the foregrounds in PAUS images have been computed using a broad-band mock catalogue interpolating the spectral energy distributions (SEDs) of objects by fitting SED templates. On top of the foregrounds, noise from any other sources (electronic, atmospheric, etc.) has been included by assuming that the sum of all noise follows a Gaussian distribution, and measuring the variance of this distribution directly from PAUS reduced images.

This chapter is divided in three sections. First, in section 4.1, we will briefly discuss the two main kinds of numerical simulations used in cosmology: n-body and hydrodynamic. In section 4.2, the three elements used for the modelled survey (hydrodynamic simulation, mock catalogue and noise) are described, and in section 4.3, we explain how these datasets are combined to simulate both PAUS images and eBOSS/DESI Lyα forest data.

4.1 Overview of cosmological simulations

In order to constrain and either verify or falsify cosmological models, the theory behind these models needs to be compared against observations. However, the Universe is a system complex enough that the only way to relate most observables to the underlying theory is by running numerical simulations based on said theoretical models.

For example, the variation of the observed 2PCF (or other clustering observables) depending on different dark matter models can only be inferred by adjusting the properties of the hypothetical dark matter and running a numerical simulation of its evolution. For smaller scales, features related to galaxy and cluster evolution rely even more on simulations to both make predictions and better understand the actual underlying physical processes.

Therefore, cosmological simulations are an invaluable tool for modern cosmology, and thanks to the continuous improvement on computational power and numerical methods, a very active research topic. We will briefly explore the two main types of numerical simulations used in cosmology: N-body and hydrodynamic. For a complete overview on the current state-of-the-art, we refer the reader to Vogelsberger et al., 2020 and the references therein.

4.1.1 N-body simulations

At cosmological scales, the dominant element that dictates both the observed structure and its evolution is dark matter. According to the current ΛCDM paradigm, cold dark matter is collisionless and only interacts through gravitational force; all the electromagnetic processes that affect baryons at macroscopic scales can be simply ignored, as well as pressure or turbulence. Consequently, simulating only the evolution of dark matter should be much less complex than taking into account its baryonic counterpart, while providing an accurate picture at large scales. This is the principle behind N-body simulations, where dark matter particles evolve as bodies interacting through gravity only.

Normally, these simulations are run by setting an initial density field at redshift high enough ($z \sim 100$), following the primordial inflationary power spectrum and the transfer function (see section 2.2.2). Dark matter particles are placed according to this initial distribution, and their evolution is computed until $z = 0$. All the information regarding all dark matter particles (position, momentum) is stored at certain time steps; these snapshots allow to recreate later the properties of the simulated Universe at different redshifts.

Computing the evolution of the dark matter particles implies solving the collisionless Boltzmann equation, which determines the evolution by of the phase-space density (i.e., the space of position and momentum where all the particles are represented as a dynamical system), and the Poisson equation, which gives the gravitational potential. In most simulations Newtonian gravity is considered, as the linear theory approximation yields the same results as general relativity, and dark matter particles do not reach relativistic speeds. Given that the simulation is confined to a finite box, periodic boundary conditions are set, which emulate the homogeneity (and the cosmological principle) observed in the real Universe.

This relative simplicity of the gravity-only evolution, together with the aforementioned approximations, make n-body simulations suitable to simulate large cosmological volumes. The first simulations already covered scales large enough to study galaxy clustering, (e.g., Aarseth, Turner, and Gott, 1979; Aarseth and Fall, 1980), but with only around ∼1000 particles per simulation (and usually, one particle assigned per galaxy). Modern state-of-the-art simulations, such Millennium XXL (Angulo et al., 2012) or Euclid Flagship (Potter, Stadel, and Teyssier, 2017) are run with $\sim 10^9 - 10^{12}$ particles in boxes with sides of the order of Gpc/h, which covers a significant portion of the sky. Zoomed simulations, which focus on much smaller volumes (of the order of ∼Mpc/h) with much higher resolution also exist, such as COCO (Hellwing et al., 2016), designed to study the halo structure and small-scale effects, or CLUES (Libeskind et al., 2009), which replicates our Local Group. Figure 4.1 shows the structure at different scales of one of the largest N-body simulations, MICE-Grand Challenge (Fosalba et al., 2013).

Given that the only output of N-body simulations are the positions of dark-matter particles, as well as its organisation in halos, filaments and other structures, some kind of model is needed to obtain derive observable properties (namely, galaxy

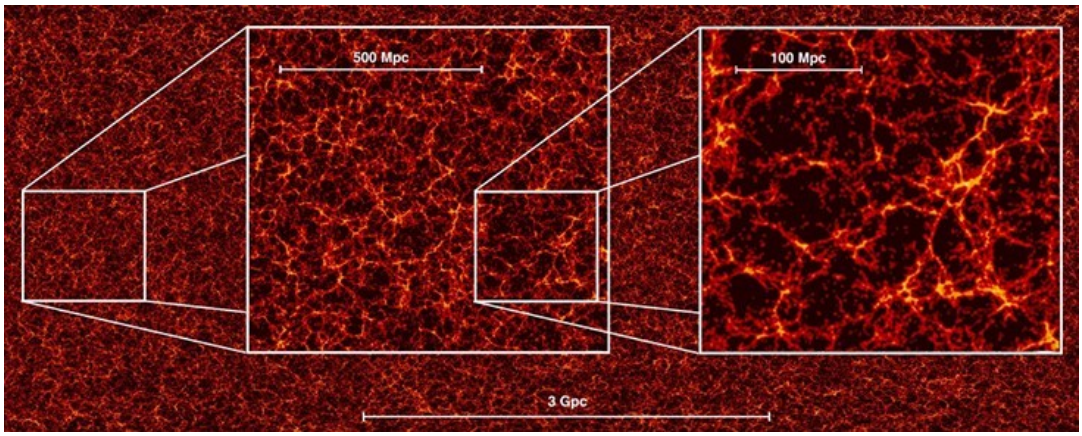

FIGURE 4.1: Example of one of the largest N-body simulations, MICE-Grand Challenge, displaying the wide dynamic range (from Gpc to Mpc scales) covered by the simulation. Extracted from Fosalba et al., 2013, Fig. 1.

positions, as well as its basic spectral properties and its magnitude) solely from the dark matter structure. Without such a model, the results of N-body simulations could hardly be compared to observations, as dark matter by itself cannot be observed. This is usually achieved with semi-analytical models (e.g., Hatton et al., 2003; Benson, 2012), which use analytical expressions to derive galaxy occupations in halos, HI mass, star formation rates, etc. from the properties of the halos and their merging trees.

Considering that the Lyα forest is a crucial element of this work, it is worth noting that one of the most remarkable successes of early cold dark matter N-body simulations was to properly emulate the absorption patterns of the Lyα forest. This finding was a decisive proof for the adoption of cold dark matter in the current standard model.

4.1.2 Hydrodynamic simulations

While N-body simulations can reproduce with a great degree of accuracy the cosmic web, baryonic distributions, as well as visible objects such as galaxies or quasars, can only be approximated via semi-analytical models (which usually are calibrated either with empirical data or other simulations, offering little to no predictive power). Besides, the effects that baryonic matter may have on the smaller structures (e.g., halo shapes), are completely neglected.

Therefore, if baryonic phenomena such as galaxy formation and evolution are to be studied, we need simulations that include baryons as another element that evolves and interacts. Such simulations are known as *hydrodynamic simulations*. Usually, baryons are included in the simulation as ideal gases with no viscosity, and after setting an initial gas distribution, they are evolved by solving the Euler equations of fluid dynamics. There are two numerical approaches to the problem: Eulerian and Lagrangian.

In Eulerian methods, the simulation volume is divided into cells, and the hydrodynamic equations are solved considering each cell of the mesh as a discrete entity, which is equivalent to formulating the equations for a static observer that does not flow with the gas particles. Given the huge dynamic range of hydrodynamic

simulations (as variables such as density or temperature can vary several orders of magnitude between a star-forming region and an underdense region of the IGM, for example), adaptive meshes are employed.

On the other hand, Lagrangian methods use as a reference discrete moving elements (sampling particles), which follow the hydrodynamic equations of motion. In this case, the equations are formulated for a moving observer that is part of the fluid, and dividing the simulation volume into a mesh is not needed. The most commonly used Lagrangian approach is the smoothed-particle hydrodynamics method, which was originally developed in the field of astrophysics (SPH, Gingold and Monaghan, 1977).

In addition to the hydrodynamics of the baryonic gas (which already are more complex than the gravitational interactions of dark matter particles), several physical phenomena regarding baryons need to be taken into account. These are usually out of the already large dynamic range of hydrodynamic simulations, and it is completely unfeasible to simulate them numerically, unlike the baryonic gas. Consequently, they are implemented with simple analytical approximations or numerical methods, known as *subresolution models* (as they emulate processes way below the numerical resolution of the simulation). Some of the most important are:

- Star formation: Stars form from the accretion of cold and dense gas. This process is impossible to replicate numerically in a cluster-sized simulation, as a galaxy of the size of the Milky Way contains $\sim 10^{11}$ stars. Usually, this is simulated by assigning a probability of generating stars to gas regions above certain density thresholds, or other gas properties. Gas particles that turn into stars are then considered collisionless, as they behave like point particles at galactic scales.

- Stellar feedback: Stars do affect their surrounding gaseous medium, either by the continuous emission of stellar winds or the drastic ejections of matter in supernovae. These processes are simulated by injecting either thermal or kinetic energy on gas particles surrounding star regions.

- AGN feedback: Active galactic nuclei (AGN) are the accreting disks surrounding supermassive black holes at the centre of galaxies. The high-energy processes in these objects result in the emission of a strong electromagnetic continuum, as well as the ejection of relativistic matter in collimated jets. Since these are well outside the dynamic range of cosmological hydrodynamic simulations, subresolution models are needed for its implementation.

Initially, hydrodynamic simulations managed to reproduce key properties of the IGM, but struggled to form realistic galaxies (e.g., much larger bulges than disks in late-type galaxies, unrealistic stellar masses). Recent simulations can produce fairly realistic late-type (spiral) and early-type (elliptical/irregular) galaxies, but there is still much to be understood about the underlying physical processes, especially subresolution models.

Given the higher complexity of hydrodynamic equations, as well as all the embedded subresolution models, hydrodynamic simulations cover much smaller volumes than their N-body counterparts; the largest hydrodynamic simulations have maximum sizes of the order of ~ 100 Mpc/h, such as EAGLE (Schaye et al., 2015) or Illustris-TNG (Springel et al., 2018). Smaller scale simulations, which focus solely on galaxy formation and evolution, without much consideration of large-scale structure, also exist, such as NIHAO (Wang et al., 2015) or Latte/FIRE (Wetzel et al.,

2016). In fig. 4.2, a example of star distribution in large-scale structure (left panel) and a simulated disk galaxy (right panel) from different Illustris-TNG simulations are shown.

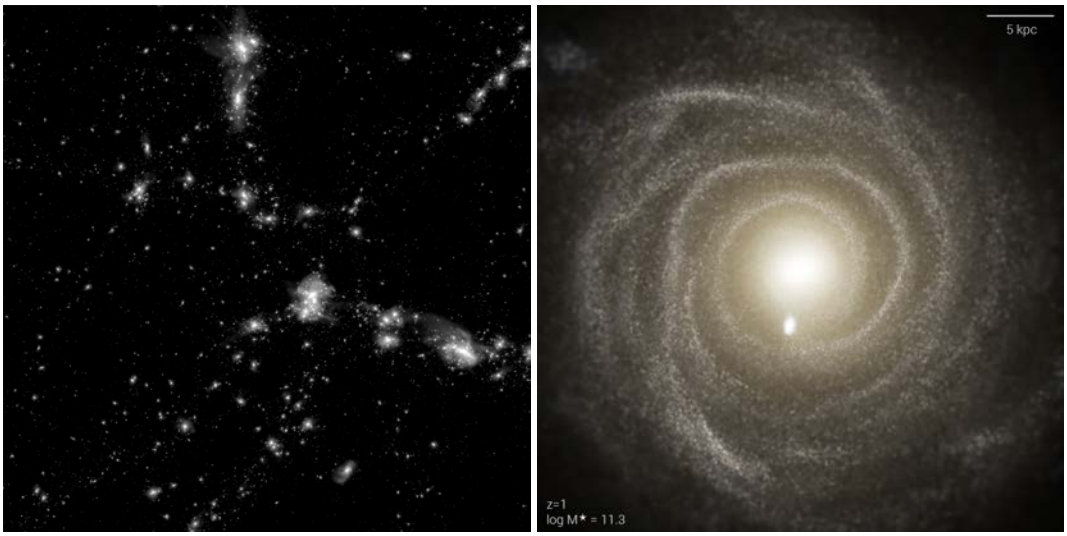

FIGURE 4.2: *Left panel:* Stellar mass distribution in a projection of Illustris TNG100-1 at $z = 0$ (cropped). *Right panel:* Detail of a large disk galaxy from Illustris TNG50 at $z = 2$. Credit for both: TNG Collaboration.

4.2 Independent simulations in this work

As we have mentioned at the beginning of the chapter, the simulated PAUS data is a combination of different cosmological simulations (hydrodynamic for Lyα data, N-body for foregrounds), plus a Gaussian realisation of atmospheric/instrumental noise. The hydrodynamic simulation is also used to model the Lyα forest data from eBOSS/DESI. In this section, we will review these cosmological simulations, as well as explain how the instrumental noise from PAUS has been estimated.

4.2.1 Hydrodynamic simulation

The hydrodynamic simulation used in this work has been performed with the P-GADGET code (Springel, 2005; Di Matteo et al., 2012), with $2 \cdot 4096^2$ particles in a 400 Mpc/h box using the cosmology specified in table 4.1. While this cosmology has small differences in parameters compared to a more standard cosmology obtained from observational surveys (e.g., table 1.1), these deviations in cosmological parameters are small enough to not affect the results of this work in any significant manner. Besides, the discrepancies in h can be entirely be accounted for with the current tensions in H_0 measurements. Therefore, for consistency the cosmology of the hydrodynamic simulation has been used for the entirety of this work.

Particle masses of $1.19 \cdot 10^7\, h^{-1} M_\odot$ and $5.92 \cdot 10^7 h^{-1} M_\odot$ were used for gas and dark matter respectively, with a gravitational force resolution of $3.25 h^{-1}$ kpc. In order to speed up the simulation, the density threshold for star formation was lower than usual, so gas particles became collisionless star particles more quickly. This density threshold was 1,000 times the mean gas density. Besides this, black hole

formation and stellar feedback were not taken into account. While this results in inaccurate stellar properties of galaxies, it does not significantly affect the IGM, and thus the simulated Lyα forest (Viel, Haehnelt, and Springel, 2004).

TABLE 4.1: Cosmological parameters of the hydrodynamic simulation; they have been also used for any other calculations carried out in this work.

Ω_Λ	Ω_m	Ω_b	h	n_s	σ_8
0.725	0.275	0.046	0.702	0.968	0.82

This simulation was originally computed for Lyα forest studies, and has already been used in several works. In Cisewski et al., 2014 and Ozbek, Croft, and Khandai, 2016, different methodologies to model the 3D IGM between Lyα forest data were tested with it, while in Croft et al., 2018 it was used to simulate Lyα IM. Figure 4.3 shows a voxel plot of the hydrodynamic simulation in both Lyα emission, in luminosity units (erg/s), and absorption, in δ flux contrast, defined as

$$\delta_\mathrm{i} \equiv \frac{e^{-\tau_\mathrm{i}}}{\langle e^{-\tau} \rangle} - 1. \tag{4.1}$$

Where τ_i is the optical depth of the Lyα forest pixel i (section 2.4.2), computed along sightlines through the simulation, as in Hernquist et al., 1996. Therefore, with this definition high values of δ correspond to regions with low HI density, and vice-versa. This δ absorption flux is expected to have a clustering bias with respect to dark matter of $b_a = 0.336 \pm 0.012$ at $z = 2.25$ (Slosar et al., 2011), including redshift distortion effects.

While the physics leading to the Lyα forest absorption are reproduced explicitly in the hydrodynamic simulation, we make predictions for the Lyα luminosity using a simple heuristic model, with an amplitude normalised using observational data; not enough is known about all sources of Lyα emission to warrant using a more detailed model.

In this model, the Lyα luminosity is proportional to the square of the baryonic density field at the scale of the spatial binning of the hydrodynamic simulation (1.56 Mpc/h). This is done with the following expression

$$L_{Ly\alpha}(r) = C_L \rho_\mathrm{b}(r)^2, \tag{4.2}$$

where $\rho_\mathrm{b}(r)$ is the baryonic density field, and C_L is a normalisation constant chosen in order to set the average Lyα luminosity density to $1.1 \cdot 10^{40}$erg/s/(Mpc)^3. This value of Lyα luminosity density is that measured from observed Lyα emitters at redshift $z = 3.1$ (Gronwall et al., 2007), which is a conservatively low value to use, as it does not include any sources of Lyα emission which are not readily observed in narrow-band Lyα surveys. This includes low surface brightness extended halos around Lyα emitters (e.g., Steidel et al., 2011) (which could host 50% or more extra Lyα luminosity density), or any other low surface brightness emission which could be difficult to detect in surveys aiming to detect objects above a threshold, but which would be included in an intensity map. The Lyα luminosity density we use can be converted to an associated star formation rate density applying a commonly used relationship between Lyα luminosity and star formation rate (SFR) of $1.1 \cdot 10^{42}$

erg/s/($M_\odot$/yr) at $z \sim 3$ (Cassata et al., 2011). This relation yields a SFR density measured from observed Lyα emitters of 0.01 $M_\odot$/yr/Mpc^3 (Gronwall et al., 2007).

Once the Lyα luminosity density is determined for a simulation cell in the model, the Lyα luminosity values are convolved with the line of sight velocity field, in order to put the Lyα emission into redshift space. This technique is similar to that used to convert the Lyα forest absorption spectra into redshift space (see e.g., Hernquist et al., 1996). The work of estimating the Lyα luminosity from the baryon distribution of the hydrodynamic simulation was mostly done by our collaborator Rupert Croft, from Carnegie Mellon University.

The baryons are unbiased with respect to dark matter, and thus in the model, the Lyα emission is expected to be biased with respect to dark matter by a factor $b_e \sim 2$ on linear scales (due to Lyα being related to the square of the baryonic density). This b_e in the model is chosen to be consistent with the measured bias of Lyα emitters at these redshifts (e.g., Gawiser et al., 2007), considering that these are the predominant sources of Lyα emission, and that the contribution of the IGM is subdominant. We note that the assumption of squaring the density will lead to a linear bias of $b_e = 2$, which may not hold at very highest densities, and this may result in artefacts in the form of extremely bright pixels. As it is explained later (section 4.3.1), a Lyα flux threshold is set for the simulated PAUS images, partially in order to account for this effect.

4.2.2 Mock catalogue/Foreground simulation

If we consider PAUS images for Lyα IM, most of the detected photons of cosmic origin will not come from Lyα at a certain redshift (depending on the filter used), but from uncorrelated sources at different redshifts than the expected Lyα emission. The main contributors to this contamination of the signal will be foregrounds, i.e., objects with lower redshift between the Lyα emission and the observer. In this work, 96.7% of all the observed flux in the simulation (averaged over all filters) was from foregrounds (after masking resolved objects).

Since the goal of this work is assessing the potential of cross-correlating PAUS with Lyα forest data, a realistic model of these foregrounds is key for our study. In order to model them, we will need a mock catalogue that spans a range of redshift large enough (at least $z = 2.75$, but ideally until $z = 6$, where the PAUS redshift range for Lyα ends), with an angular size comparable to the Lyα forest/emission simulation box. Besides, all objects in the catalogue must have their observed SEDs in the PAUS wavelength range (455-855 nm) and with resolution higher than PAUS FWHM ($\Delta\lambda < 13$ nm).

The two first requirements (redshift range and angular size) are met by already available mock catalogues generated with N-body simulations and semi-analytical models, but none of them contain direct SED information. Such mock catalogues are intended to reproduce large surveys, with the only spectral information available being either broad bands, which do not meet the resolution requirement, or emission lines, which are insufficient to generate the foregrounds.

Our approach to this problem is to take a mock catalogue with broad bands, and interpolate SEDs for all objects by fitting SED templates to the broad bands. The mock catalogue selected is a lightcone (e.g., a larger simulation in a certain line of sight, constructed by combining the same N-body simulation at different redshifts) originally developed to simulate data from the Euclid satellite, made from a run of the Millennium Simulation using WMAP7 cosmology (Guo et al., 2013). This lightcone is complete up to magnitude 27 in Euclid H band, which makes it ideal for

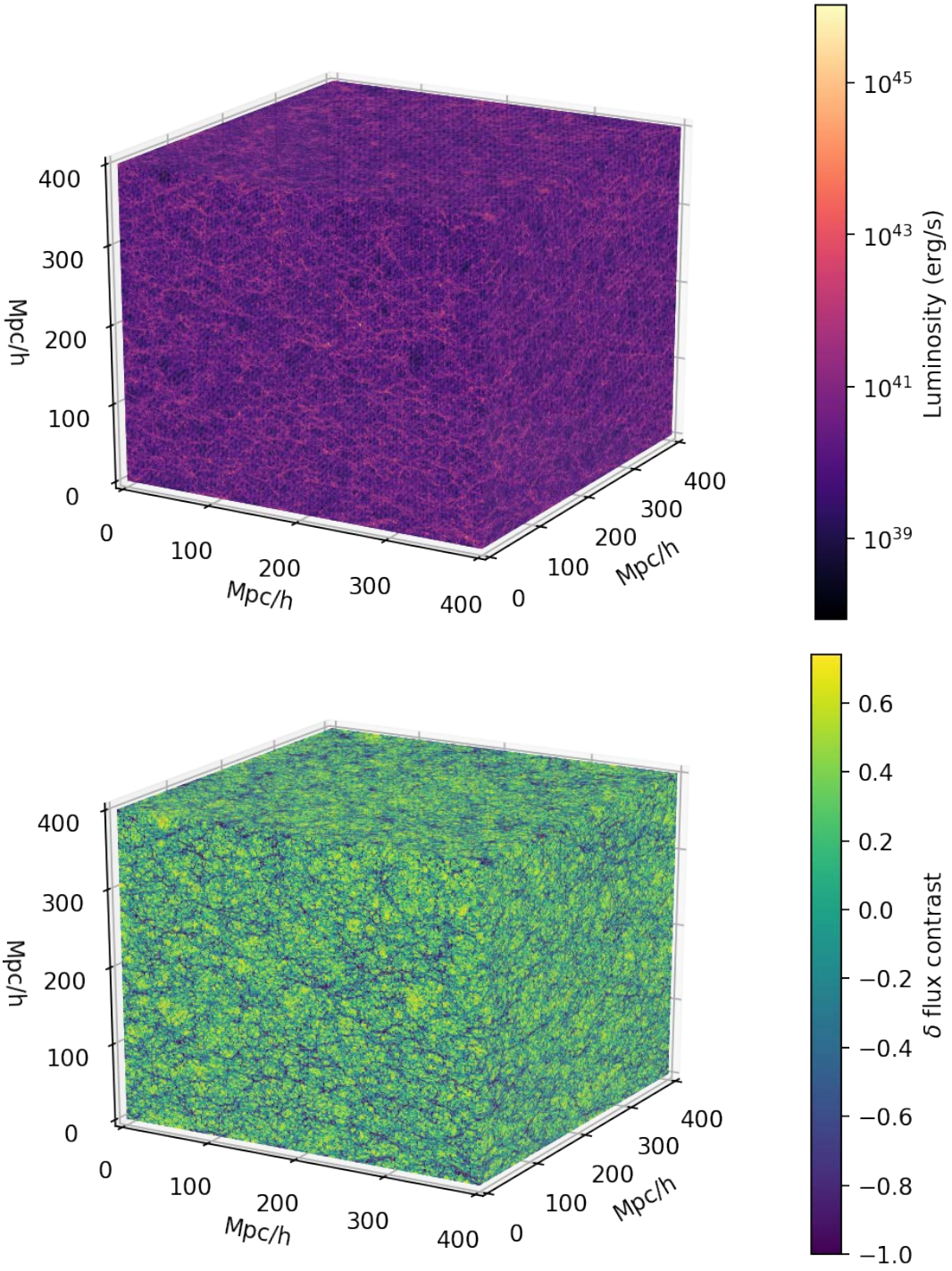

FIGURE 4.3: Hydrodynamic simulation used for this work. *Top panel:* Extended Lyα emission, in erg/s and logarithmic colour scale. *Bottom panel:* δ flux contrast, used to model the Lyα forest.

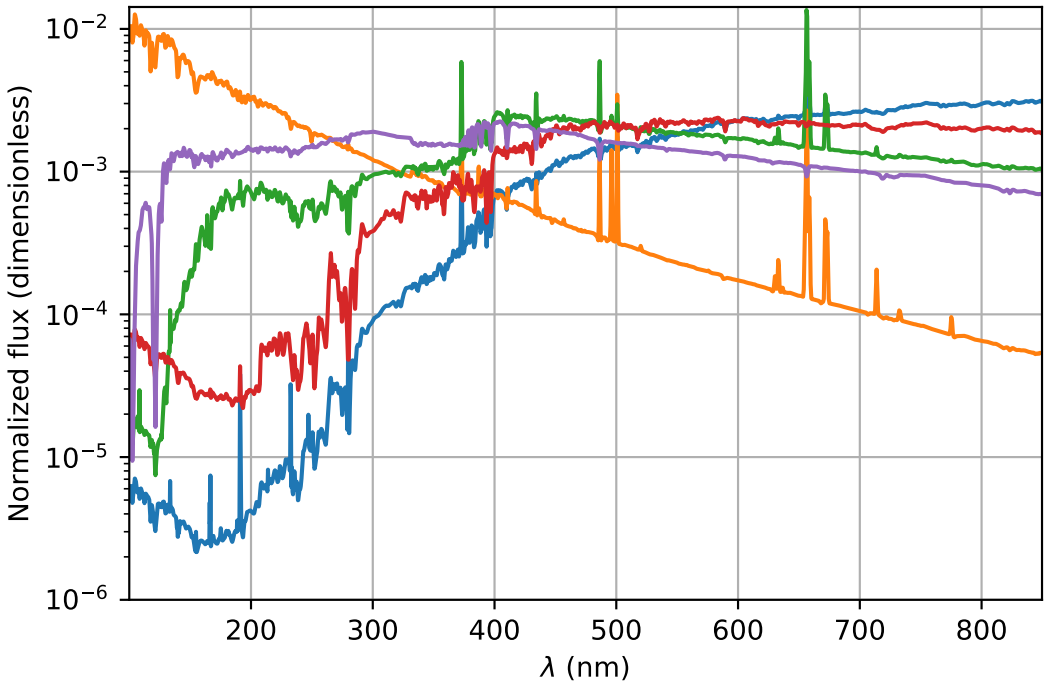

FIGURE 4.4: The five SED templates used for foreground simulation, normalised to facilitate visual comparison.

foreground simulations (since most mock catalogues do not reach such depths). The semi-analytical model applied to compute galaxies is GALFORM (Gonzalez-perez et al., 2014), and the lightcone was constructed with the technique described in Merson et al., 2013.

In order to interpolate SEDs, we have considered the SDSS *ugriz* bands (fig. 3.3, black lines) from this mock and the five SED templates defined by Blanton and Roweis, 2006, which we show in fig. 4.4. For the five templates, their *ugriz* band values have been computed in a fine redshift grid ($\Delta z \sim 0.01$). These template bands are used as the elements of a coordinate basis, and for any object the coefficients of the linear combination of templates that gives the *ugriz* bands of the object can be computed with the following expression

$$\begin{pmatrix} u_{\rm obj} \\ g_{\rm obj} \\ r_{\rm obj} \\ i_{\rm obj} \\ z_{\rm obj} \end{pmatrix}_{z} = \begin{pmatrix} u_1 & g_1 & r_1 & i_1 & z_1 \\ u_2 & g_2 & r_2 & i_2 & z_2 \\ u_3 & g_3 & r_3 & i_3 & z_3 \\ u_4 & g_4 & r_4 & i_4 & z_4 \\ u_5 & g_5 & r_5 & i_5 & z_5 \end{pmatrix}_{z_{\rm grid} \sim z} \times \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}. \tag{4.3}$$

Where the left hand array are the bands of the object *obj* at redshift z, the right hand array X are the coefficients of the linear combination of the templates, and the matrix is the basis of template bands at the redshift $z_{\rm grid}$ closest to the redshift of the object z. This is a simple linear system that has a single exact solution as long as the basis matrix is invertible (which has been checked for all $z_{\rm grid}$). However, the coefficients X must be all non-negative for the SED to make physical sense (since the SED templates are patterns of emitted flux for galaxies, and thus subtracting them has no physical meaning). Therefore, instead of finding the analytical solution, the

coefficients are computed using non-negative least squares. This numerical method is approximate, but on average yields relative errors of a few percent when recovering the original bands. Once these coefficients are obtained, the linear combination of SED templates using the coefficients is computed for all objects, thus generating a full mock catalogue with high spectral resolution SEDs.

4.2.3 PAUS Noise

In addition to the foregrounds, PAUS images have noise from a large variety of sources (electronic, airglow, etc.), together with the intrinsic variability between nights (seeing, moonlight, etc). Instead of simulating each one of these components with a physical model, we have measured them directly from PAUS images. For each one of the 7 filters considered, 8 exposures in 10 different pointings in the COSMOS field have been blindly selected as a representative sample to evaluate the noise. All of these images were already reduced by the PAUS pipeline, but some additional processing was carried out to emulate the additional reduction that would be necessary for IM applications.

First, resolved sources were removed by applying a sigma-clipping filter with 3σ threshold in 5 iterations; ideally, the masks could be extracted from a reference catalogue, but as a preliminary study sigma-clipping is enough to virtually remove all resolved objects. The masked pixels were replaced by random values drawn from a Gaussian distribution with the same mean and σ as the unmasked pixels of the image, to avoid having empty pixels that would result in an overestimation of σ (since σ needs to be computed for the pixel size of the simulation, which is larger than the CCD pixel size, masked pixels would result in artificially smaller samples inside a simulation pixel, and thus a higher σ). Figure 4.5 shows one of the reduced science image from PAUS used for the noise measurement (top panel), and the result after applying sigma-clipping (bottom panel).

Second, once resolved objects were masked, the median flux value was computed and subtracted for all the images. This was done to cancel out the variability in sky brightness due to moon phases and time of observation, which may modify the average background flux by a factor of few. While this erases all Lyα clustering signal at scales larger than the CCD ($\sim$12 Mpc/h in its smallest dimension at $z = 3$), this approach is enough for noise determination in this preliminary work. A proper modelisation of the moonlight and sky brightness as a function of date and time could remove this variability without erasing the large-scale Lyα signal, but it is out of the scope of this thesis. However, it is a pending task if Lyα IM is to be performed on PAUS data (or other optical imaging surveys).

After this processing, σ could be measured directly from the resulting images, but it would include not only the electronic and atmospheric noise aforementioned, but also the variance due to the cosmic foregrounds and the Lyα signal, which are already considered in our simulation.

In order to remove all signals from cosmic origin and keep only the electronic and atmospheric noise, we have stacked all exposures for all pointings, but applying an scaling factor of -1 to half of them. Since the number of exposures is even, any signal that should remain constant between exposures would tend to zero (e.g., both Lyα and foreground emission), while the variability due to atmospheric and electronic components remains. This has been done using SWARP (Bertin et al., 2002), disabling background removal (to not artificially decrease the resulting noise), and cropping the regions of the stacked image where there was not a full overlap of all the 8 exposures. Figure 4.6 show an example of the resulting stacked image.

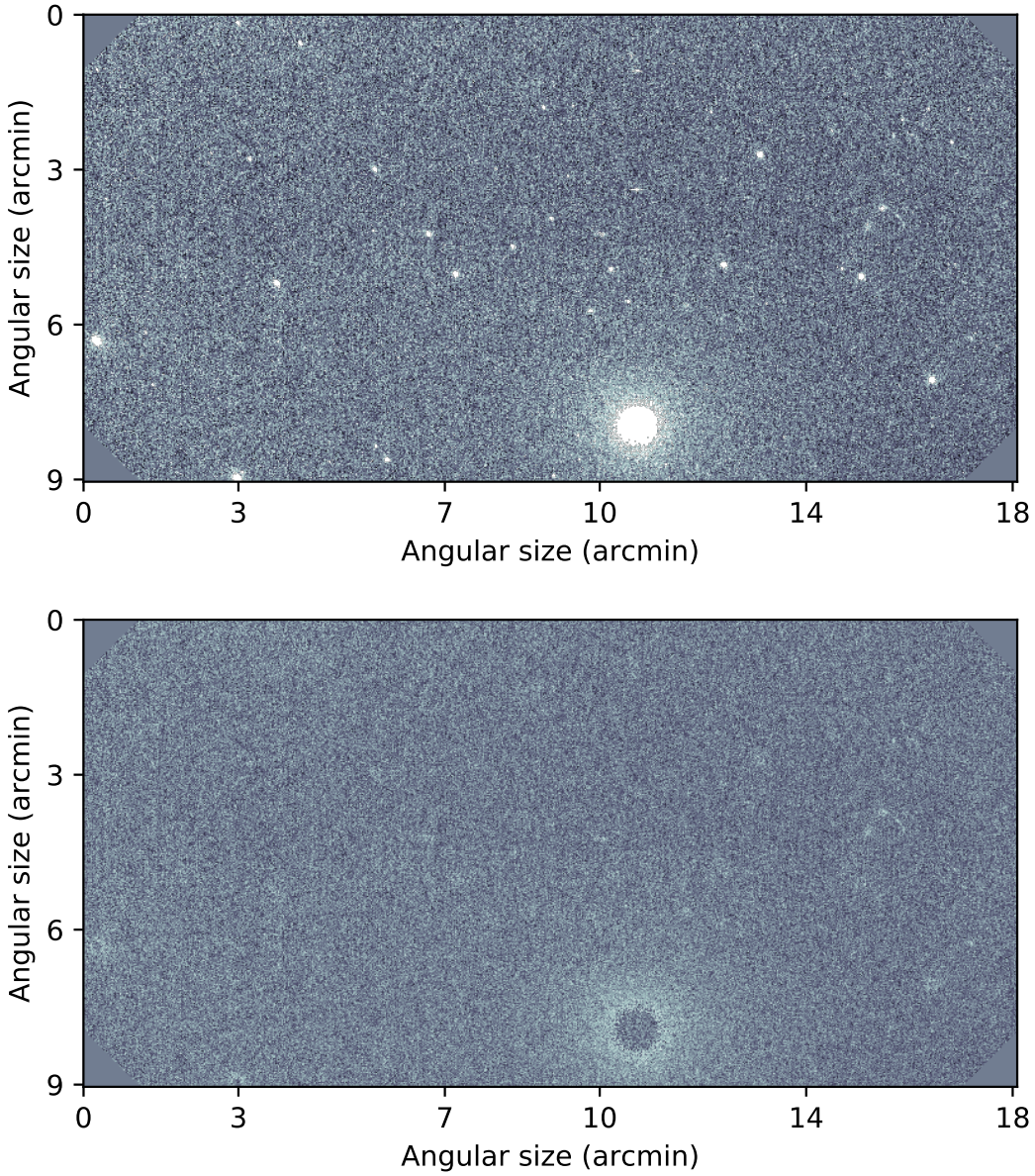

FIGURE 4.5: Example of PAUS image in the 455 nm filter used to measure noise. *Top panel:* Reduced science image. *Bottom panel:* The same image, after sigma-clipping and replacement of masked pixels by Gaussian values.

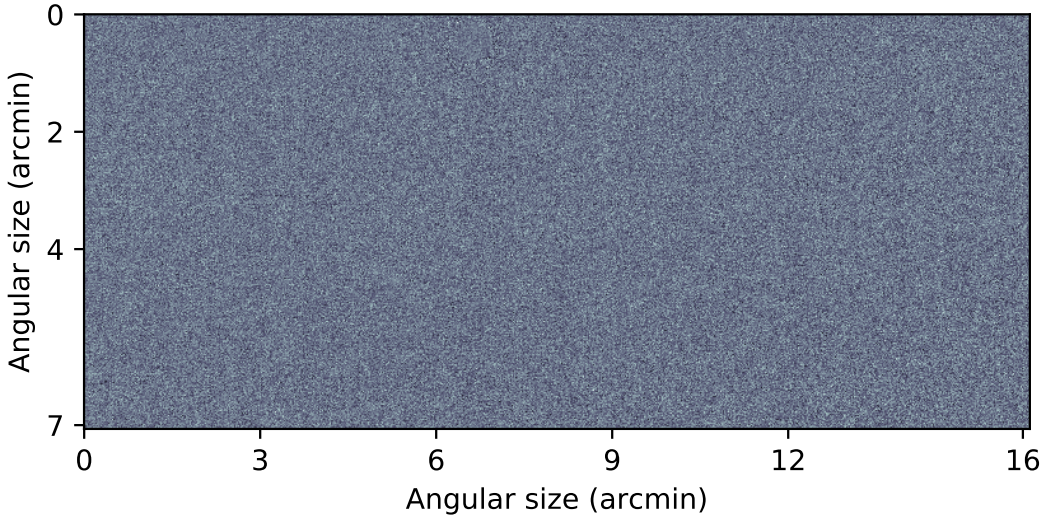

FIGURE 4.6: Stacking of 8 clipped exposures in the 455 nm filter, with alternate sign. Same pointing as the images in fig. 4.5.

For each one of these stacked images, the σ was computed for increasing pixel sizes, starting by the intrinsic pixel size of the CCD, and going above the pixel size of our hydrodynamic simulation. The flux values of these increasing pixel sizes were computed by adding the values of all pixels inside them instead of averaging, since the hydrodynamic simulation considers the total Lyα luminosity in each 3D pixel, not its spatial average. Given that the images are stacked, the resulting σ has been divided by $\sqrt{N_{\rm exp}}$, to scale the result to a single exposure.

Figure 4.7 shows the average σ for each filter versus pixel size; the vertical line represents the pixel size of the simulation, and the dashed line an extrapolation of the σ vs pixel size considering uncorrelated noise (averaged for the 7 filters). This extrapolation has been determined with

$$\sigma_1 = \frac{\theta_1}{\theta_0}\sigma_0, \tag{4.4}$$

where σ_0 and θ_0 are the standard deviation and angular size for the original pixels of the image, and σ_1, θ_1 its counterparts for the new pixels. This expression comes from the fact that the sum of uncorrelated Gaussian variables has a σ^2 equal to the sum of all the σ^2 of the individual Gaussian distributions. In this approximation, we consider each pixel to be an uncorrelated Gaussian with equal $\sigma_{\rm pixel}$; thus, the sum of n pixels will have $\sigma_{\rm sum} = \sqrt{n}\sigma_{\rm pixel}$. Since we are adding CCD pixels to form larger pixels where σ is computed, this $\sqrt{n}$ factor will be equal to the ratio of angular sizes, which results in eq. (4.4).

By looking at fig. 4.7, it is clear that the σ measured from the images displays a noticeable correlation for pixel scales larger than 30 arcsec, since it shows a much steeper slope in logarithmic scale. With the approach we have followed, it is certain that this correlation is not of cosmic origin, but other than that, we can not speculate more on the causes for this observed correlation, which are left for future research.

Therefore, we have considered two different noise levels: first, the measured σ for the simulation pixel size, which would be the intersection of the coloured lines with the vertical black line in fig. 4.7, and the uncorrelated σ following eq. (4.4).

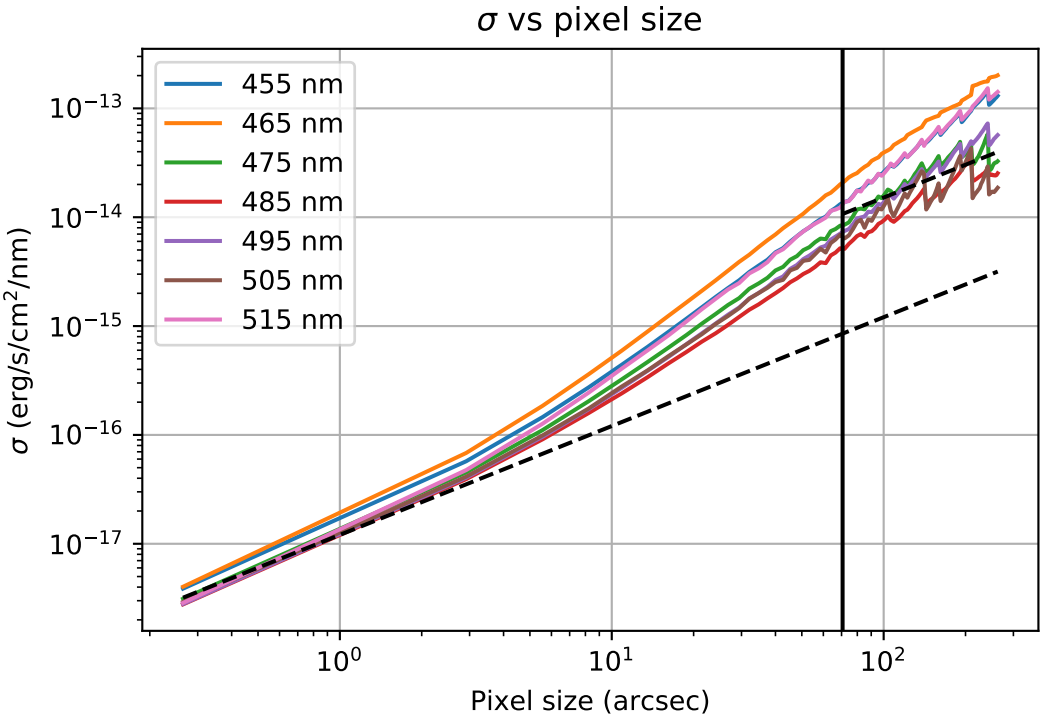

FIGURE 4.7: Average measured σ of masked images vs pixel size, for the seven filters used in this work. The dashed line shows the extrapolated mean for uncorrelated noise, and the vertical line the pixel size of the simulation.

TABLE 4.2: σ_{noise} measured for the pixel size of the simulation, in (erg/s/cm^2/nm)$\cdot 10^{-15}$, for the seven narrow-band filters, as well as its value scaled for three exposures, σ_{3exp}, and 18 exposures, σ_{18exp}.

λ (nm)	455	465	475	485	495	505	515
σ_{noise}	13.52	20.78	8.46	5.15	7.29	6.62	13.14
σ_{3exp}	7.80	12.00	4.89	2.97	4.21	3.82	7.58
σ_{18exp}	3.19	4.90	1.99	1.21	1.72	1.56	3.10

TABLE 4.3: σ_{noise} extrapolated as uncorrelated noise to the pixel size of the simulation, in (erg/s/cm^2/nm)$\cdot 10^{-16}$, for the seven narrow-band filters, as well as its value scaled for three exposures, σ_{3exp}, and 18 exposures, σ_{18exp}.

λ (nm)	455	465	475	485	495	505	515
σ_{noise}	10.32	10.74	8.32	7.39	7.46	7.79	7.63
σ_{3exp}	5.96	6.20	4.81	4.27	4.31	4.50	4.40
σ_{18exp}	2.43	2.53	1.96	1.74	1.76	1.84	1.80

The former represents the most realistic case if the cross-correlation with actual data was to be computed now, while the latter is an hypothetical case where through further work on image reduction all noise correlations are removed, and only the uncorrelated and irreducible electronic noise remains.

In table 4.2, the mean σ_{noise} for each filter, measured at the pixel size of the simulation, as well as the scaled noise for three exposures σ_{3exp}, is shown, by dividing σ_{noise} by a factor of $\sqrt{N_{exp}}$. A hypothetical case for a deeper PAUS (complete up to $i_{AB} < 24$) is also considered, since it is a possibility currently being explored. This would imply multiplying by six the current exposure time for all survey pointings, hence the σ_{18exp}. Table 4.3 shows the same data for the uncorrelated noise approximation. Overall, the correlation in the noise increases σ_{noise} by a factor of $\sim$10.

4.3 Simulation of PAUS Lyα IM

All the elements described in the previous section (hydrodynamic simulation, foregrounds from mock catalogue and PAUS noise) need to be properly combined in order to obtain a realistic simulation of PAUS images for IM. Besides, Lyα absorption data from the hydrodynamic simulation needs some processing to emulate the space sampling of the Lyα forest; these procedures will be explained in this section.

4.3.1 PAUS images: Lyα emission

In order to simulate the PAUS images for the cross-correlation, the elements explained in the previous subsection (Lyα emission from the hydrodynamic simulation, foregrounds from the mock catalogue and Gaussian noise) must be converted to units of observed flux density (erg/s/cm^2/nm) and merged into the seven narrow-band filters.

Since the hydrodynamic simulation gives Lyα emission in luminosity units (erg/s), the first step is to compute the comoving coordinates of all pixels of the simulation from the point of view of the observer. Assuming the cosmology of the simulation, and knowing that the simulation snapshot is at $z = 3$, we consider the comoving distance from the observer to the centre of the box to be the radial comoving distance at

redshift 3, $\chi(z=3)$. Knowing this, the comoving coordinates of all cells of the simulation with respect to the observer are also known (as well as their edges), assuming that the three axes of the simulation box are RA, dec and radial directions respectively. The bins of the hydrodynamic simulation are not in spherical coordinates but Cartesian, however, given the small angular size of the sample, the small-angle approximation can be applied.

With the comoving radial distance of all cells known, and the relation $\chi(z)$ given by the cosmology, the inverse relation $z(\chi)$ can be computed numerically, and thus a redshift can be assigned to each cell. This allows to compute the luminosity distance simply with its definition for a flat cosmology (section 1.2.7)

$$d_L(z) = (1+z)\cdot\chi(z). \tag{4.5}$$

Moreover, given that all the emitted flux is Lyα, the rest frame wavelength is also known ($\lambda_{\mathrm{Lyff}} = 121.567$ nm), which yields the observed wavelength range of all cells in the hydrodynamic simulation, and thus all redshift bins (following the small angle approximation, all cells in the same radial distance bin will have the same redshift, and thus observed wavelength range). With all these elements computed, the observed flux density for all PAUS cells comes from the following expression

$$f_{\lambda\, i} = \frac{L}{4\pi {d_L(z_{\mathrm{i}})}^2 \Delta\lambda_{\mathrm{i}}^{\mathrm{obs}}}. \tag{4.6}$$

Where L is the cell luminosity given by the hydrodynamic simulation (erg/s), d_L the luminosity distance in cm, and $\Delta\lambda_{\mathrm{i}}^{\mathrm{obs}}$ the observed wavelength range for the redshift bin of the cell, all corresponding to the PAUS cell i.

Having computed the observed flux density for all PAUS cells, the redshift bins of the PAUS simulation need to be merged to simulate the wavelength bins given by PAUS filters. In order to do so, PAUS filters are considered to have top-hat response functions 10 nm wide, ranging from 455 nm (bluest filter) to 845 nm (reddest). Following this criterion, the redshift bins of the simulation completely fill the seven bluest filters, which also limits the cross-correlation to seven filters in this work. The last four redshift bins of the simulation fall outside the seventh filter; these bins are discarded for the simulation of PAUS images. For each one of the simulated PAUS filters, all the redshift bins of the hydrodynamic simulation that fall inside the wavelength range of the filter are merged into a single one, with its flux value being the mean of the merged bins (since observed fluxes are the average flux density over the response function).

With redshift bins already merged to simulate PAUS filters, the average Lyα redshift for each filter can be used to convert from observed flux densities (erg/s/cm^2/nm) to absolute flux densities (erg/s/nm), with the following expression

$$F_{\lambda\, i} = 4\pi {d_L(z_{\mathrm{nb}})}^2 f_{\lambda\, i}. \tag{4.7}$$

Where z_{nb} is the redshift of Lyα in the respective narrow band. This is done in order to cancel out the dimming of observed Lyα flux with redshift (due simply to the increasing distance between said emission and the observer), which would introduce an artificial gradient in the emission field to be cross-correlated. However, the previous conversion to observed fluxes was necessary, since we can only convert to absolute fluxes with observational data using the observed redshift, i.e., PAUS redshift bins, not the much finer redshift bins of the original simulation.

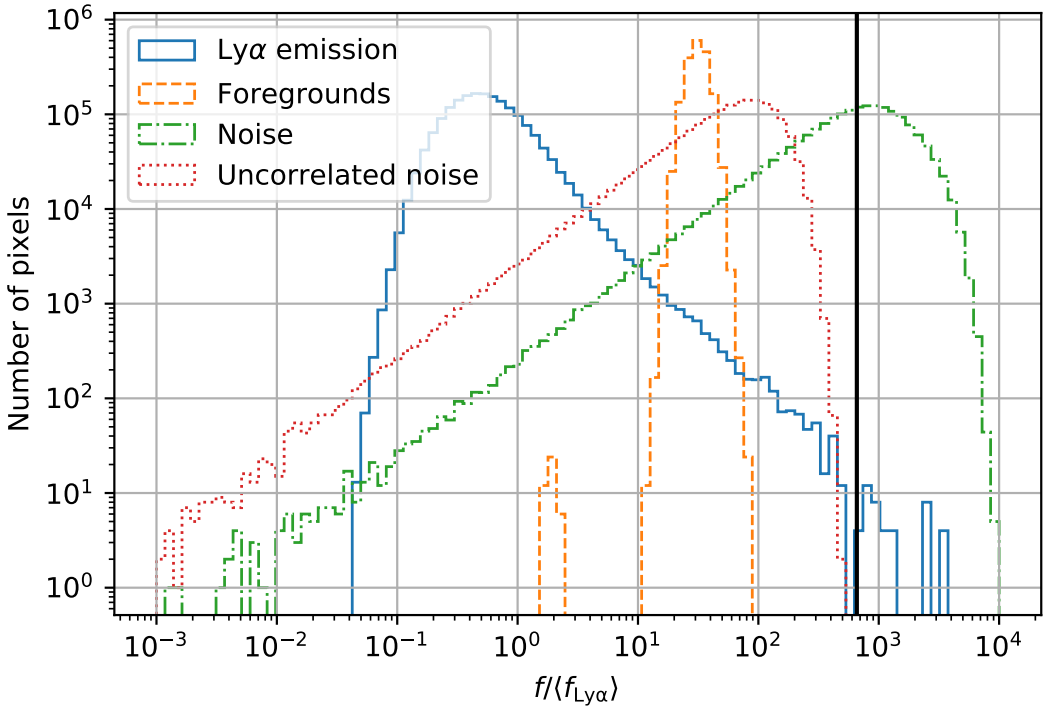

FIGURE 4.8: Logarithmic histograms of the ratio between fluxes and mean Lyα fluxes, for the Lyα emission, foreground emission, and instrumental noise, both the measured noise and the uncorrelated extrapolation (in absolute value). Noise values for the current PAUS case, $\sigma_{3\text{exp abs}}$. The vertical line represents the imposed Lyα threshold.

On top of this conversion to absolute fluxes, a realistic threshold can be imposed to Lyα fluxes, both to remove possible artefacts that may be derived from the assumption that Lyα luminosity is proportional to baryon density squared, and also to account for the fact that resolved objects will be removed from PAUS images before cross correlating (which may remove some bright Lyα emitters at high redshift).

The chosen Lyα absolute flux threshold is 10 times the brightest pixel of the simulated foregrounds, whose computation will be explained in section 4.3.2. This value is chosen assuming that the foreground simulation gives a realistic estimate of how much unresolved flux can be expected, and taking into account that resolved objects are masked based on their g band luminosity. This broad band has FWHM=138.7 nm (fig. 3.3), which is one order of magnitude wider than PAUS narrow bands. Therefore, Lyα emission observed in a PAUS filter will be reduced by a factor of 10 when observed in the g filter. A maximum value of of $1.53 \cdot 10^{-5}$ erg/s/nm was set as a threshold, which affected only 0.0024 % of all pixels. To visualise the extent of this threshold, fig. 4.8 shows histograms of absolute fluxes for the Lyα emission, foregrounds and instrumental noise, divided by the mean Lyα flux and together with the Lyα threshold, represented as a vertical line.

After all these steps, the result is a simulation of Lyα extended emission in PAUS filters. However, given the redshift and the size of the simulation, it only covers $\sim$ 25 deg^2, with an angular pixel size of 1.38 arcmin^2; since the expected area to cross-correlate is 100 deg^2, the simulation is replicated four times in mosaic pattern, which effectively covers the expected area. The result can be seen in fig. 4.9.

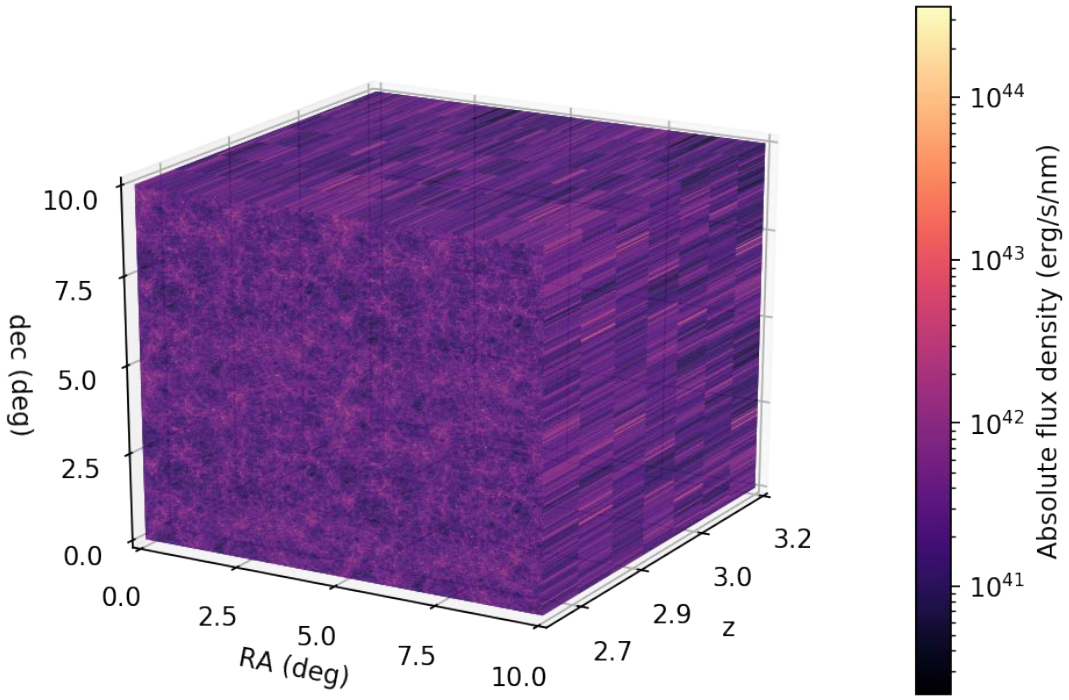

FIGURE 4.9: Simulation of the observed Lyα flux in PAUS images, in comoving coordinates according to Lyα observed redshift. Absolute flux densities in erg/s/nm.

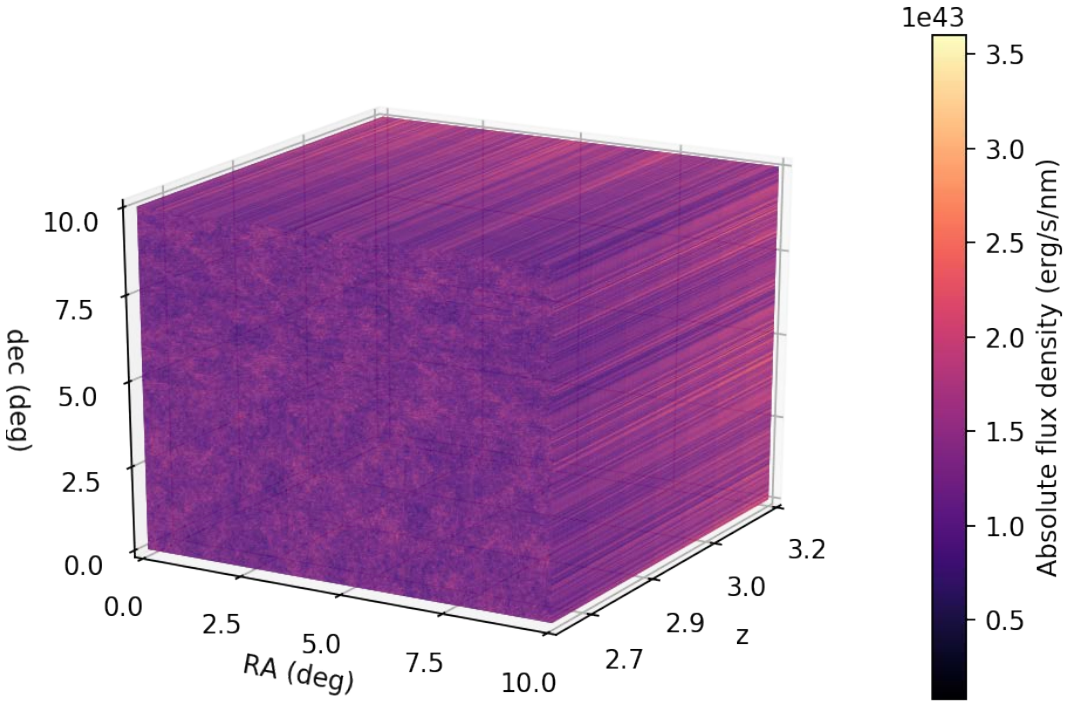

FIGURE 4.10: Simulation of the observed Lyα flux in PAUS images with foregrounds emission included, in comoving coordinates according to Lyα observed redshift. Absolute flux densities in erg/s/nm.

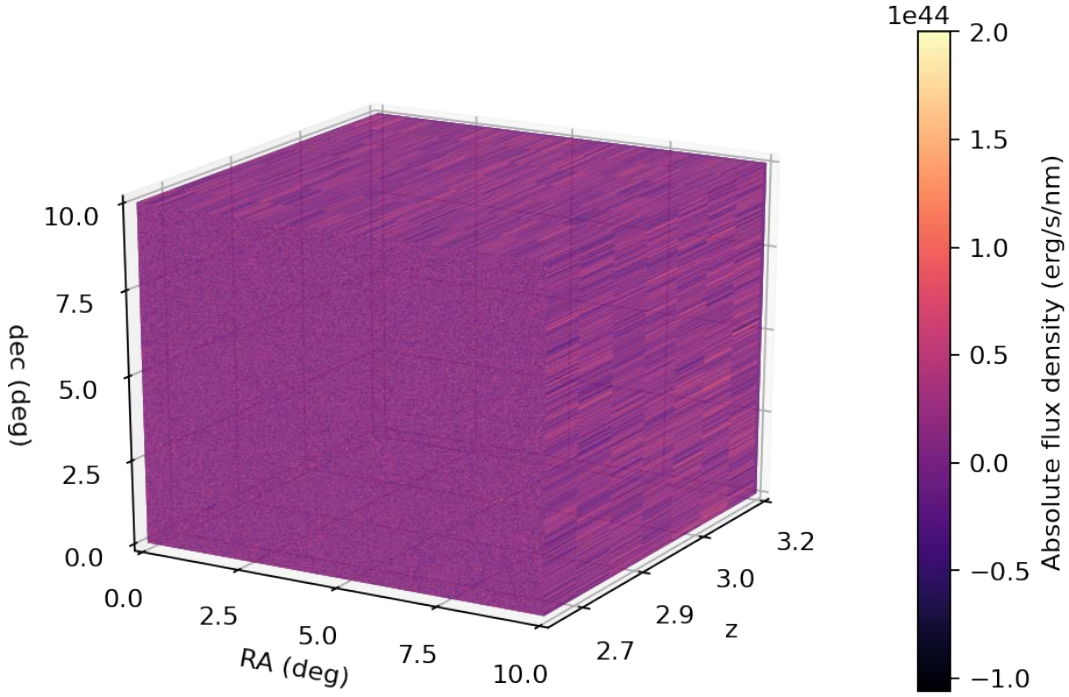

FIGURE 4.11: Simulation of the observed Lyα flux in PAUS images with foregrounds emission and instrumental/atmospheric noise included, in comoving coordinates according to Lyα observed redshift. Absolute flux densities in erg/s/nm. Instrumental/atmospheric noise considered for the current PAUS case, $\sigma_{3\text{exp abs}}$, using the uncorrelated noise approximation.

4.3.2 PAUS images: foregrounds

Given that resolved objects will be removed from PAUS images before cross-correlating, only the objects too faint to be resolved must be included in the foreground simulation. The PAUS reference catalogue is complete up to magnitude 25 in the g band; consequently, only the objects in the mock lightcone dimmer than this value are selected. Besides this, since the lightcone is elliptical in angular coordinates, it is cropped to the largest inscribed rectangle. This rectangle is smaller than the 25 deg^2 at z$\sim$3 of the hydrodynamic simulation, so it is repeated in a mosaic pattern and cropped to cover the same angular area as the original Lyα simulation.

All the foreground objects have their SEDs computed by template fitting, as explained in the previous section, and they are binned in RA and dec using the same angular bins as the Lyα flux simulation. Since the templates are fitted to apparent magnitudes, by using the definition of AB magnitude the interpolated SEDs are already in observed flux units of erg/s/cm^2/nm.

For each one of these RA x dec pixels, the net observed SED is computed as the sum of the SEDs inside the bin. These stacked SEDs are then integrated and averaged over the response functions of the seven PAUS blue filters according to the expression below, which gives the observed foreground flux,

$$f_{\rm nb} = \frac{\int_0^\infty d\lambda f_{\rm SED}(\lambda) R_{\rm nb}(\lambda)}{\int_0^\infty d\lambda R_{\rm nb}(\lambda)}. \tag{4.8}$$

Here $f_{\rm SED}$ is the flux density of the interpolated SED, $R_{\rm nb}$ the response function of a certain narrow band, and $f_{\rm nb}$ the observed flux density in that narrow band. With this expression the observed foreground flux in the PAUS filters is obtained; in order to convert to absolute fluxes eq. (4.7) is used.

The result is a three-dimensional array covering $\sim$25 deg^2 that can be directly added to the Lyα observed flux simulation. As in the Lyα flux case, this array needs to be replicated four times in a mosaic pattern for an effective coverage of 100 deg^2. This time, however, for each replication the array is rotated clockwise (keeping the redshift direction the same), in order to ensure that each 25 deg^2 subset is a different realisation of Lyα emission+foregrounds (if the rotation was not performed, replicating the arrays for a 100 deg^2 would be analogous to sampling the same 25 deg^2 area four times). The result of these simulated foregrounds can be seen in fig. 4.10.

This rotation introduces discontinuities in the foreground structure, since the periodic boundary conditions of the mock catalogue are broken. Nevertheless, the cross-correlation is computed by selecting cubes of PAUS cells around forest cells, so only forest cells close enough to the discontinuities will be affected by them.

As shown in fig. 5.3, the cross-correlation is only computed in perpendicular (angular) direction up to 20 Mpc/h. Given that the whole angular size of the simulation is 800 Mpc/h, and that the discontinuities are two straight lines dividing the simulation in RA and dec, this leaves <10% of the forest cells potentially affected by the discontinuities. Also, the dominant noise contribution is instrumental noise, not the foregrounds, so even in the small fraction of forest cells affected by discontinuities the effects of these on the cross-correlation should be fairly small.

4.3.3 PAUS images: Combination and noise

Considering that both have the same units and the same binning, the Lyα and foregrounds absolute flux simulations can be directly added into a total absolute flux

TABLE 4.4: $\sigma_{\text{noise abs}}$ with noise correlation, in (erg/s/nm)·10^{44}, for the seven narrow-band filters, as well as its value scaled for three exposures, $\sigma_{3\text{exp abs}}$, and 18 exposures, $\sigma_{18\text{exp abs}}$.

λ (nm)	455	465	475	485	495	505	515
$\sigma_{\text{noise abs}}$	9.12	15.05	6.56	4.27	6.45	6.24	13.16
$\sigma_{3\text{exp abs}}$	5.27	8.69	3.79	2.47	3.72	3.60	7.60
$\sigma_{18\text{exp abs}}$	2.15	3.55	1.55	1.01	1.52	1.47	3.10

TABLE 4.5: $\sigma_{\text{noise abs}}$ following the uncorrelated extrapolation, in (erg/s/nm)·10^{43}, for the seven narrow-band filters, as well as its value scaled for three exposures, $\sigma_{3\text{exp abs}}$, and 18 exposures, $\sigma_{18\text{exp abs}}$.

λ (nm)	455	465	475	485	495	505	515
$\sigma_{\text{noise abs}}$	6.96	7.78	6.46	6.13	6.60	7.34	7.64
$\sigma_{3\text{exp abs}}$	4.02	4.49	3.73	3.54	3.81	4.24	4.41
$\sigma_{18\text{exp abs}}$	1.64	1.83	1.52	1.44	1.56	1.73	1.80

array. The only step left to properly simulate PAUS observations is to add the instrumental and atmospheric noise. For this simulation, we have modeled this noise as a Gaussian distribution of mean zero and σ dependent on the filter. This σ is the instrumental noise directly measured from images and scaled for the number of exposures, as specified in table 4.2, converted to absolute flux units according to eq. (4.7). Two cases have been considered: the σ measured at the pixel size of the simulations (table 4.2), and the extrapolation considering uncorrelated noise (table 4.3), which yields lower values of σ.

These absolute flux noise values, $\sigma_{\text{noise abs}}$, as well as the scaled value that is used, $\sigma_{3\text{exp abs}}$, and the hypothetical deep PAUS, $\sigma_{18\text{exp abs}}$, are displayed in table 4.4 for the real noise case, and in table 4.5 for the uncorrelated extrapolation. The final result of Lyα flux+foregrounds+instrumental noise is shown in fig. 4.11. Only the uncorrelated noise case is shown, as the resulting figure in this case is already noise-dominated.

With this simulation, despite repeating both the Lyα emission and the foregrounds in a mosaic pattern, we ensure that the cross-correlation always samples a different combination of signal+noise, since instrumental noise is generated for the full simulation and foregrounds are rotated.

While it may be argued that the clustering signal from Lyα emission is repeated, the only caveat of this is that cosmic variance may be underestimated. Given that the original diameter of the hydrodynamic simulation is 400 Mpc/h, far above the homogeneity scale (e.g., Gonçalves et al., 2018), and that the predominant sources of noise are by far foregrounds and instrumental noise (as seen in section 6.2), any effect cosmic variance may have on the result is negligible.

4.3.4 eBOSS/DESI: Lyα forest

To simulate the Lyα forest data of eBOSS/DESI surveys, the hydrodynamic absorption simulation show in fig. 4.3 is replicated four times in a mosaic pattern, as if it was shown in the PAUS simulation.

After this operation, random cells in the simulation array are selected with the quasar density redshift distribution shown in fig. 3.7 (depending on the survey to be

simulated), with the redshift of each cell computed as for the Lyα emission simulation. The RA and dec coordinates of the quasar cells are selected randomly from a uniform distribution. The total number of quasar cells (i.e., the number of quasars in the sample) is also computed from the redshift distribution, considering that the simulation has an angular area of 100 deg^2 and that only quasars with $z > 2.7$ are to be included (since quasars at lower redshift will have all Lyα forests outside the redshift range of the simulation).

The cells between the quasar cells and the observer (the cells in the same angular bins and negative redshift direction) are considered Lyα forest cells, including the quasar cells themselves. Only these forest cells are taken into account for cross-correlation; everything else in the hydrodynamic simulation is masked.

In addition to this, if a quasar is at redshift high enough so that Lyβ forest appears at $z > 2.7$, its forest cells that would be covered by the Lyβ region are also masked, given that these regions of the quasar spectrum contain both Lyα and Lyβ absorption lines superimposed from different redshifts. While these Lyβ forest regions can be used for cross-correlation studies (e.g., Blomqvist et al., 2019), here we adopt the conservative approach and remove them from the cross-correlation. These masked Lyβ cells account for 12% of the total forest cells.

Regarding the SNR of the forest data, we take as a reference the mean SNR values displayed in Chabanier et al., 2019, Table 2. These correspond to a high quality sample of the first eBOSS release, and thus are an optimistic estimate of what can be expected in both future eBOSS releases and DESI. For the redshift bin closer to our study ($z \sim 3$), the eBOSS data shows $\langle \mathrm{SNR} \rangle = 6.5$ per forest pixel. However, this value needs to be scaled to the bin sizes of our forest simulation with the following expression,

$$\mathrm{SNR}_{\mathrm{simulation}}(\lambda) = \sqrt{\frac{\Delta\lambda_{\mathrm{simulation}}}{\lambda \langle R_{\mathrm{eBOSS}} \rangle^{-1}}} \langle \mathrm{SNR}_{\mathrm{eBOSS}} \rangle, \tag{4.9}$$

where $\Delta\lambda_{\mathrm{simulation}}$ is the wavelength bin size for our forest simulation (determined as in section 4.3), and $\langle R_{\mathrm{eBOSS}} \rangle$ the mean resolution of eBOSS ($\sim$4350 according to the data reduction in Chabanier et al., 2019). This results in a higher SNR, that increases linearly from 9.97 at the lowest redshift to 10.29 at the high redshift end. Gaussian noise is added to each forest cell according to the determined SNR.

A voxel representation of this Lyα forest simulation, displaying only forest cells used for cross-correlation, is shown in fig. 4.12, both for eBOSS and DESI expected quasar densities.

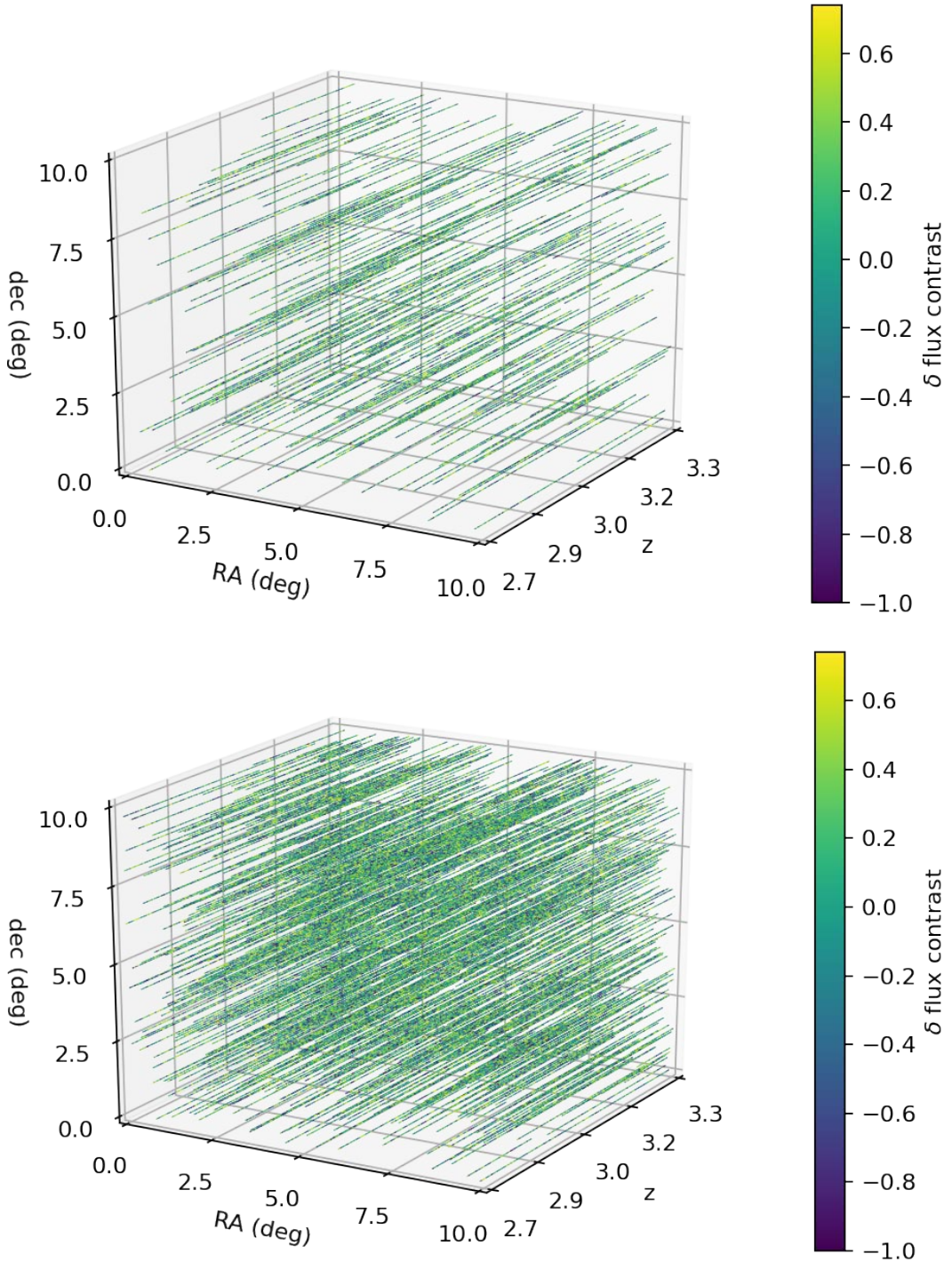

FIGURE 4.12: 3D simulations of the Lyα forest sampled pixels. *Top panel:* eBOSS. *Bottom panel:* DESI.

Chapter 5

Computation of the cross-correlation

As we have already discussed, the goal of this thesis is to simulate Lyα IM by cross-correlating PAUS with the Lyα forest data from eBOSS/DESI. So far, we have already explained how the observational data has been simulated in the last chapter, as well as laying out the theoretical basis needed to properly follow the work in the chapters before.

Therefore, the only topic left before discussing the results is *how* the 2PCF is computed; this is the subject of this chapter, divided in three subsection. First, in section 5.1 the mathematical estimator used to compute the cross-correlation will be explained, as well the error estimation and the biases derived from foregrounds and other sources of noise. Then, in section 5.2, the code elaborated as a part of this thesis to compute said cross-correlation will be briefly described, with a focus on the algorithm structure. Finally, in section 5.3 the theoretical model of the 2PCF, needed to validate and interpret the results, is discussed.

5.1 Cross-correlation estimator

5.1.1 Estimator definition

In order to compute the cross-correlation from the PAUS and eBOSS/DESI simulated datasets, an estimator of the 2PCF is needed, i.e., a function applicable to a real sample that tends to yield the real value of the underlying statistical distribution. We have already discussed the theoretical definition of the 2PCF in section 2.1.3, but strictly speaking, it can only be applied to smooth scalar fields completely defined over $\mathbb{R}^n$ (which is never going to be the case with observational data).

Generally, the 2PCF estimators used in cosmology are designed to correlate pairs of discrete points, as this is the product of galaxy surveys (e.g., see Kerscher, Szapudi, and Szalay, 2000). The main challenge these estimators face is to minimise the biases introduced by finite boundary conditions and irregular survey volumes. However, in this work the elements to cross-correlate are uniform discrete divisions of the sample volume (as it is usually the case with IM).

Besides, the boundary of the simulation is regular with no incomplete regions, and the distances sampled by the cross correlation are one order of magnitude smaller than the simulation size (35 Mpc/h vs 400 Mpc/h, see fig. 5.3). Therefore, any boundary effects are very likely to be small, especially when compared to the dominant noise components in PAUS images (both foregrounds and instrumental or atmospheric). Consequently, a more straightforward estimator has been chosen,

$$\hat{\xi}(r_n) \equiv \frac{\sum_{\mathrm{i}} \left(\delta_{\mathrm{i}} \sum_{\in \mathrm{Bin}(\mathrm{r_n})} \phi_{\mathrm{j}} \right)}{\sum_{\mathrm{i}} \left(1 \sum_{\in \mathrm{Bin}(\mathrm{r_n})} 1 \right)}. \tag{5.1}$$

This estimator is defined for distance bins r_n. Since the cells to be cross-correlated have finite volumes, distances are assumed from the coordinates of their centres. Regarding the other terms in the equation, δ_{i} is the δ flux of the forest cell i, as defined in eq. (4.1), and ϕ_{j} is the absolute flux contrast for the pixel j in simulated PAUS images, defined as

$$\phi_{\mathrm{j}} \equiv \frac{F_{\lambda\, j}}{\langle F_\lambda \rangle} - 1. \tag{5.2}$$

In other words, this estimator is the average value of the products of all cell pairs in a certain distance bin, so it can be interpreted as a discrete equivalent of the theoretical definition in eq. (2.3). Normally, the average computed by this estimator is weighted by a function of the pipeline error, as well as additional errors terms derived from data reduction (e.g. Font-Ribera et al. (2012)). However, for this preliminary work the error in simulated PAUS images is approximately constant, with only slight variations between filters (see Table 4.4), and the Lyα forest error has been considered negligible, so no weighting has been applied.

This distance r in eq. (5.1) is defined as the total distance between cells (monopole cross-correlation), but it could also be defined as the distance projected onto the line of sight (parallel cross-correlation, $\xi(r_\parallel)$), or perpendicular to it (perpendicular cross-correlation, $\xi(r_\perp)$). Consequently, the parallel and perpendicular estimators $\xi(r_{\parallel\,\mathrm{n}})$ and $\xi(r_{\perp\,\mathrm{n}})$ can be defined simply by switching the definition of distance, $|i-j|$, by $|i-j| \cdot \vec{u}_{\mathrm{los}}$ and $|i-j| \times \vec{u}_{\mathrm{los}}$ respectively (where $\vec{u}_{\mathrm{los}}$ is the unit vector parallel to the line of sight). All these three kinds of 2PCF will be considered in this work.

5.1.2 Jackknife error

If the 2PCF values obtained from the estimator are to hold any scientific validity, they need to be accompanied by their error values (i.e., the σ of the estimator). Generally, the simplest approach would be to apply the estimator to different observed samples, and compute the error as the σ of the values of the estimator for all samples.

However, in cosmology we only have one Universe to observe, and thus repeating the experiment to obtain a different set of observations is out of the question. Consequently, one common approach is the use of resampling techniques, which give an estimation of the error in any statistic from a single observational sample. Even when working solely with simulations, as in this case, resampling techniques are still applied, as cosmological simulations are computationally very expensive.

For this work, the error of the estimator has been computed jackknife resampling, a resampling method widely used in cosmology, which has been largely proven to yield reliable error estimations (Escoffier et al., 2016; O'Connell et al., 2019). In order to apply this technique, first the simulated volume need to be divided in N equal subsamples. Then, N values of the estimator are computed, each one determined over all the observed volume but the subsample k. Following the notation in eq. (5.1), this can be expressed as

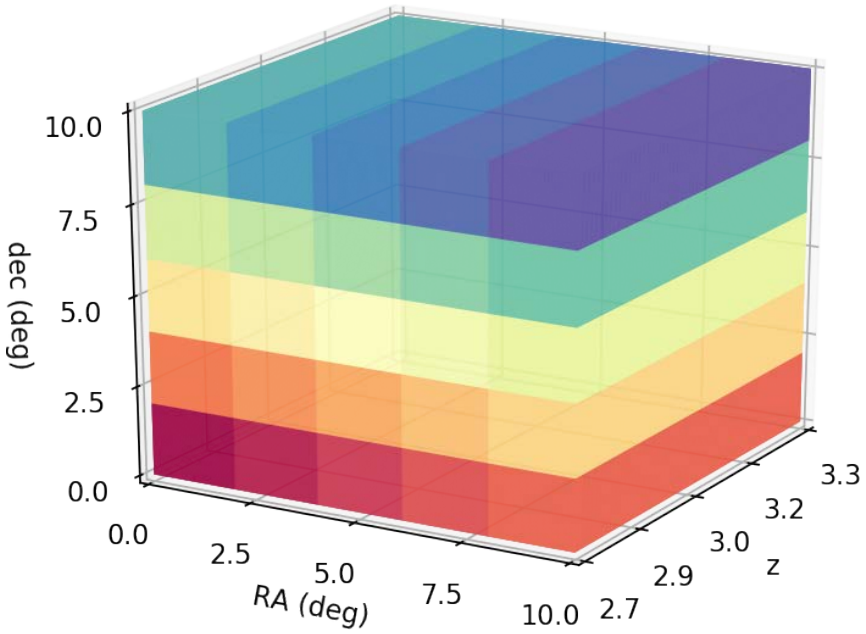

FIGURE 5.1: Jackknife division of the simulated volume (both for PAUS and eBOSS/DESI). Each colour represents a jackknife subsample.

$$\hat{\xi}_{\rm k}(r_{\rm n}) \equiv \frac{\sum_{{\rm i}\notin {\rm V_k}} \left(\delta_{\rm i} \sum_{\in {\rm Bin(r_n)}} \phi_{\rm j}\right)}{\sum_{{\rm i}\notin {\rm V_k}} \left(1 \sum_{\in {\rm Bin(r_n)}} 1\right)}, \tag{5.3}$$

where $V_{\rm k}$ is the designated jackknife subsample. For our particular case, this means that for $\hat{\xi}_{\rm k}$ all Lyα forest cells are considered but the ones lying inside the subvolume $V_{\rm k}$. Once all the $\hat{\xi}_{\rm k}(r_n)$ have been computed, the error of the estimator is determined as

$$\hat{\sigma}_{\xi}(r_{\rm n}) = \sqrt{\sum_{\rm k}^{N} \left[\hat{\xi}_{\rm k}(r_{\rm n}) - \hat{\xi}(r_{\rm n})\right]^2}. \tag{5.4}$$

For this work, the simulation has been divided in 25 subsamples by imposing uniform cuts in RA and dec. Since space is not sampled uniformly in redshift in this cross-correlation (because Lyα forest available data depends on the quasar redshift distributions), no cuts have been performed in redshift, so all jackknife subsamples cover the whole redshift range of the simulation. Figure 5.1 displays the jackknife divisions in the simulated volume.

5.1.3 Noise bias

The cross-correlation estimator introduced in eq. (5.1) is biased if at least one of the signals being cross-correlated contains noise of mean different than zero, which is of particular importance for this study. In order to demonstrate this, let us assume that

the estimator is used to cross-correlate two arbitrary observable scalar fields, $f(\mathbf{r})$ and $g(\mathbf{r})$. For both fields, a finite number of samples at different points are observed, f_i and g_i, and from these points the respective means $\langle f \rangle$ and $\langle g \rangle$ are computed. In order to apply the estimator, the contrasts of both fields need to be determined, which, as in eq. (4.1) and eq. (5.2), would be done with the following expressions

$$f_{\text{contrast i}} = \frac{f_i - \langle f \rangle}{\langle f \rangle}; \qquad g_{\text{contrast i}} = \frac{g_i - \langle g \rangle}{\langle g \rangle}. \tag{5.5}$$

If δ_i and ϕ_j are replaced in eq. (5.1) by $f_{\text{contrast i}}$ and $g_{\text{contrast j}}$, and these are substituted by its definition in eq. (5.5), the following expression can be obtained

$$\xi(r) = \frac{\sum_i \left[(f_i - \langle f \rangle) \sum_j^r (g_j - \langle g \rangle) \right]}{\langle f \rangle \langle g \rangle \sum_i \left(1 \sum_j^r 1 \right)}. \tag{5.6}$$

Here the second summation in the right side of eq. (5.1) has been rewritten as $\sum_j^r$ for simplicity. Now, let us consider that the field $g(\mathbf{r})$ is the sum of two independent fields, the signal $S(\mathbf{r})$ and the noise $N(\mathbf{r})$, so

$$g(\mathbf{r}) = S(\mathbf{r}) + N(\mathbf{r}). \tag{5.7}$$

By our definition, the noise $N(\mathbf{r})$ is uncorrelated with $f(\mathbf{r})$, so for a sample large enough a hypothetical estimated cross-correlation between $f(\mathbf{r})$ and $N(\mathbf{r})$ would tend to zero. Following eq. (5.6), this can be expressed as

$$\sum_i \left[(f_i - \langle f \rangle) \sum_j^r (N_j - \langle N \rangle) \right] \to 0. \tag{5.8}$$

Conversely, the hypothetical cross-correlation $\xi_S(r)$ between $f(\mathbf{r})$ and $S(\mathbf{r})$ would be

$$\xi_S(r) = \frac{\sum_i \left[(f_i - \langle f \rangle) \sum_j^r (S_j - \langle S \rangle) \right]}{\langle f \rangle \langle S \rangle \sum_i \left(1 \sum_j^r 1 \right)}. \tag{5.9}$$

Nevertheless, only the field $g(\mathbf{r})$ can be observed, and thus the only cross-correlation that can be computed is that of the $f(\mathbf{r})$ with $S(\mathbf{r})$ plus $N(\mathbf{r})$:

$$\xi_{S+N}(r) = \frac{\sum_i \left\{ (f_i - \langle f \rangle) \left[\sum_j (S_j - \langle S \rangle) + \sum_j (N_j - \langle N \rangle) \right] \right\}}{\langle f \rangle \langle S + N \rangle \sum_i \left(1 \sum_j 1 \right)}. \tag{5.10}$$

If a sample large enough is assumed, eq. (5.8) holds true, and since the noise component of the cross-correlation tends to zero, the denominator in eq. (5.9) and eq. (5.10) is identical. Therefore, the following relation can be derived between the hypothetical cross-correlation of the signal, $\xi_S(r)$, and the actual cross-correlation of the signal with noise, $\xi_{S+N}(r)$,

$$\xi_{S+N}(r) = \frac{\langle S \rangle}{\langle S + N \rangle} \xi_S(r). \tag{5.11}$$

If the noise of the observable $g(\mathbf{r})$ had mean zero, we would have $\xi_{S+N}(r) = \xi_S(r)$, and thus the estimator would be unbiased. However if we consider PAUS images to be the observable $g(\mathbf{r})$, the noise $N(\mathbf{r})$ would be the foregrounds plus

instrumental noise. The first component necessarily has a mean larger than zero, since it is a sum of observed fluxes, while the second also should in principle, given that it includes effects such as scattered light and airglow, which are strictly positive.

Nevertheless, this noise bias does not affect the SNR, and thus the probability of detection, which is the main result presented in chapter 6. Considering that the error is computed via jackknife resampling (i.e., the σ of the cross-correlation computed for different subsamples), this noise bias will multiply the cross-correlation value and its error equally, and therefore will cancel out when computing the SNR.

5.2 Correlation code

When a cosmological study requires computing correlations between samples (either observed or simulated), most times a public, already existing code is used. These codes are specifically developed and maintained for cosmology, and the bulk of the codes is written in low-level languages such as C. This allows for a high degree of memory management and optimisation, at the expense of higher development and maintenance costs compared to high-level languages commonly used in scientific computing, such as Python.

Some examples of these correlation codes for cosmology would be TREECORR (Jarvis, Bernstein, and Jain, 2004), a highly versatile code capable of computing the 2PCF and 3PCF of galaxy positions, lensing measurements, or scalar fields (e.g., CMB), CORRFUNC (Sinha and Garrison, 2020), focused on the 2PCF of galaxy pairs, or POLSPICE (Szapudi et al., 2001), developed exclusively for CMB correlations.

However, the characteristics of the Lyα IM simulated in this work differ greatly from the most frequent cross-correlations computed in cosmology: the Lyα forest is a high-resolution tracer that samples the space only in the line-of-sight dimension, while the Lyα emission in PAUS images can be interpreted as a scalar field in fine redshift bins. Consequently, a cross-correlation code has been developed from scratch for this work. In this section we will discuss its main aspects as an algorithm; the actual core of the correlation code, together with a more technical explanation of its variables, is displayed in appendix A.

Given that the cross-correlation code has to be developed in the framework of a PhD thesis, with the sole aim of simulating the PAUS-eBOSS/DESI cross-correlations, the chosen language for the code has been Python; a more efficient with a lower-level language would be out of the scope of the work. Regardless, computing times are small enough to not require further optimisation in our case, as it will be shown later.

The aim of the code is relatively simple: to compute eq. (5.1) for a given set of distance bins (either monopole, parallel or perpendicular). The simplest approach would be to compute all possible pairs between PAUS pixels (cells of the hydrodynamic Lyα emission simulation) and possible forest pixels (cells of the hydrodynamic Lyα absorption simulation), then select all the cell pairs of interest for each distance bin. This is a pure brute-force approach (since all the possible combinations are computed), and while simple and reliable, it is also very inefficient. Therefore, to optimise the code until reasonable computation times are achieved, some features that separate it from a pure brute-force approach have been implemented.

First, the Lyα forest cells are only a very small fraction of the hydrodynamic absorption simulation, since most of the IGM is not sampled by quasars. It is mandatory then to compute only the pairs with absorption cells that are forest cells, as

everything else (which is the great majority of the pairs) will be automatically discarded. This has been achieved by generating a data array containing only the forest data, and iterating over it (first over every forest, then over every cell of the respective forest). Algorithm 1 is a pseudo-code describing the cross-correlation algorithm; these two iterations can be seen after the necessary initialisations.

Second, even if we only consider the forest cells, there is no point in computing the pairs with all possible PAUS cells. The hydrodynamic simulation has a diameter of 400 Mpc/h, and if we consider that the actual PAUS simulation repeats it in a mosaic pattern, the area covered spans a length of $\sim$800 Mpc/h in RA and dec directions. The scale of homogeneity is $<$200 Mpc/h for the largest estimates, if the criterion for homogeneity is the 2PCF tending to zero. Consequently, cell pairs at larger distances will yield no information. Besides, to compute the 2PCF reliably at a certain distance a volume of diameter approximately one order of magnitude larger that this distance is needed; if the sample volume is smaller, the intrinsic variance of the structure makes the estimation unreliable. Hence, our correlation is limited to $r < 40$ Mpc/h, which is only a small fraction of PAUS cells lying in the vicinity of the considered forest cell, compared to the whole simulation.

Therefore, for a given forest cell, only a cube of surrounding PAUS cells is selected, and the product of the forest cell with these PAUS cells is determined (as well as their distances, which are pre-computed at the beginning of each forest with the small-angle approximation). The size of the cube (e.g., the number of adjacent PAUS cells in both line of sight and RA/dec directions) is an entry parameter of the code, and must be chosen accordingly with the distance binnings for r, $r_\parallel$ and $r_\perp$ (so no distance bins are left without pairs, nor pairs are computed outside of all distance bins). The computation of distances between the forest cell and the surrounding cube of PAUS cells is displayed in algorithm 1 at the beginning of the iteration over forests, while the selection of PAUS cells appears at the beginning of the nested loop over forest cells.

Third, a more technical aspect of the algorithm is the use of external libraries when possible, instead of Python code. Like any other Python script for scientific computation, we make extensive use of NumPy for the management of data arrays, as well as a myriad of functions and operators that can be executed over said arrays. These NumPy libraries are made with highly-optimised C code and vectorised (i.e., designed to work with a whole array of elements at once), and thus are orders of magnitude faster than a function for the same purpose written directly in Python.

Consequently, the bulk of operations most be handled exclusively by NumPy functions, leaving Python as little work as possible. This is achieved by minimising the number of Python loops, especially if they are already nested into other loops. In order to do so, the whole array of elements to be iterated through is passed to the NumPy function; since this function is vectorised, it is significantly faster than operating over every single element at once with a Python loop. The pseudo-code in algorithm 1 shows the exact loop structure of the actual Python script: cross-correlation is computed just with two levels of nesting (one for each forest, and one for each cell in the forest). The instructions of the *if* statement inside the second loop are only executed when the memory limit is reached, and thus are ignored for the vast majority of iterations.

Once the rationale behind the code design has been explained, we can describe in detail the structure of algorithm 1. First, the pseudo-code shown there is just the core of the cross-correlation script; it is assumed that the following elements are already defined:

Algorithm 1: Algorithm of the cross-correlation code.

```
   /* Initialisation of arrays to store histogram results          */
1  for i in jackknife indices do
2      Cell pair values for each monopole bin (dd_mono_i, 1D);
3      Number of cells pair for each monopole bin (np_mono_i, 1D);
4      Cell pair values for each parallel and perpendicular bin (dd_2d_i, 2D);
5      Number of cell pairs for each parallel and perpendicular bin (np_2d_i, 2D);
   /* Initialisation of cell pair data arrays, limited by memory   */
6  Cell products (cell_products);
7  Line of sight (LoS), RA and dec distances (los_dist, ra_dist, dec_dist);
8  Jackknife indices (jk_ind);
   /* Iteration over all forests                                   */
9  for forest in forests do
10     Compute unit vectors in LoS, RA and dec directions;
11     Compute relative distances to all PAUS cell pairs;
       /* Iteration over all forest cells                          */
12     for forest cell in forest do
13         if inside same PAUS cell as previous forest cell then
14             Keep the same cube of PAUS cells;
15         else
16             Determine the new cube of PAUS cells centred at the forest cell;
17             Crop the cells of the cube that lie outside the survey volume;
18         Multiply cells in the PAUS cube by the forest cell, store in cell_values.;
19         Store LoS, RA and dec distances in los_dist, ra_dist, dec_dist;
20         Store jackknife indices (from forest cell) for each cell pair in jk_ind;
           /* Computing histograms if memory full/last iteration   */
21         if memory limit reached or last iteration then
22             Compute perpendicular distance as
                dist_perp=(dist_ra^2+dist_dec^2)^{1/2};
23             for i in jackknife indices do
24                 Select elements in cell pair arrays with jk_ind= i;
25                 Compute histogram binning by dist_mono, no weighting. Add
                    result np_mono_i;
26                 Repeat same histogram, weighting by cell_pairs. Add result to
                    dd_mono_i;
27                 Compute 2D histogram binning by , no weighting. Add result
                    to np_2d_i;
28                 Repeat same 2D histogram, weighting by cell_pairs. Add result
                    to dd_2d_i;
29             Reinitialise cell_pairs, los_dist, ra_dist, dec_dist_i, jk_ind;
30 for i in jackknife indices do
31     Compute \hat{\xi}(r)_i = \sum_{j\neq i} dd\_mono_j / \sum_{j\neq i} np\_mono_j;
32     Compute \hat{\xi}(r_\parallel, r_\perp)_i = \sum_{j\neq i} dd\_2d_j / \sum_{j\neq i} np\_2d_j;
   /* N_jk is the number of jackknife subsamples                   */
33 Compute monopole as \hat{\xi}(r) = \sum_i \hat{\xi}(r)_i / N_{jk};
34 Compute monopole error following eq. (5.4);
35 Compute 2D cross-correlation as \hat{\xi}(r_\parallel, r_\perp) = \sum_i \hat{\xi}(r_\parallel, r_\perp)_i / N_{jk};
36 Compute 2D cross-correlation error following eq. (5.4);
37 Average \hat{\xi}(r_\parallel, r_\perp) over r_\perp to obtain \hat{\xi}(r_\parallel). Do the same for its error;
38 Average \hat{\xi}(r_\parallel, r_\perp) over r_\parallel to obtain \hat{\xi}(r_\perp). Do the same for its error;
```

- For the Lyα forest:
  - A 2D array (one dimension for forests, another for cells in each forest) with δ values.
  - An analogous 2D array with jackknife indices (integers which indicate to what jackknife subsample each forest cell belongs to).
  - A 3D array of Cartesian coordinates of each cell (one dimension for coordinates in RA, dec and z directions, two for forests and forest cells).
- For the simulated PAUS images:
  - A 3D array (since the whole volume is sampled) with δ values.
  - A 4D array with their Cartesian coordinates (one dimension for coordinates in RA, dec and z directions, three to point to each cell of the array).

In addition to this, the distance binnings in r, $r_\parallel$ and $r_\perp$ must already have been defined as entry parameters, as well as the proper size of the cube of PAUS cells whose pairs are to be computed. With all the data necessary, the bulk of the cross-correlation is a combination of two tasks: computing the products of cell pairs, and classifying the pair counts and products in the proper redshift bins. The algorithm tackles both separately to make the most out of NumPy functions.

First, the arrays necessary to store the histogram results are initialised from line 1 to line 5. Two 2PCFs are considered: the monopole (binned in r) and the 2D 2PCF (binned in both $r_\parallel$ and $r_\perp$, from which the parallel and perpendicular 2PCF are computed). For each one, two arrays are initialised: one for the number of cell pairs per bin (histogram without weighting), and another for the sum of all cell pair products of the respective bin (histogram weighted by the cell pair products). Dividing the second by the first is akin to eq. (5.1), which gives the estimated 2PCF. Given that the error is computed via jackknife, these values of pairs and total cell product values per bin need to be stored separately for each jackknife subsample, which gives an extra dimension to these arrays.

Second, the arrays to store all the cell pairs data have to be defined, which is done from line 6 to line 8. These are large 1D arrays, whose size is computed according to the memory allocation given as an entry parameter. These arrays are the input data to compute the histograms from, and since NumPy functions are vectorised (including the calculation of histograms), the larger the array the faster the calculation (up to a certain limit, see fig. 5.2 and the subsequent discussion), so their size will be as large as the memory allocation permits. The following arrays are defined:

- One for the values of the cell products.
- Three for distances of cell pairs in line of sight, RA and dec directions.
- One to store the jackknife indices of the forest cell of each pair. This will not be a direct input to the histogram functions, but a mask that allows to quickly select all pairs corresponding to a given jackknife subsample.

With all the initialisations done, these cell pair data arrays are filled from line 9 to line 20; once these arrays are full (more precisely, once they would be go over its limit in the next iteration), or the loop has gone through all of the forest cells, histograms are computed from to line 21 to line 29. Finally, from line 30 to line 38 the 2PCFs are computed, following the estimator described in section 5.1.

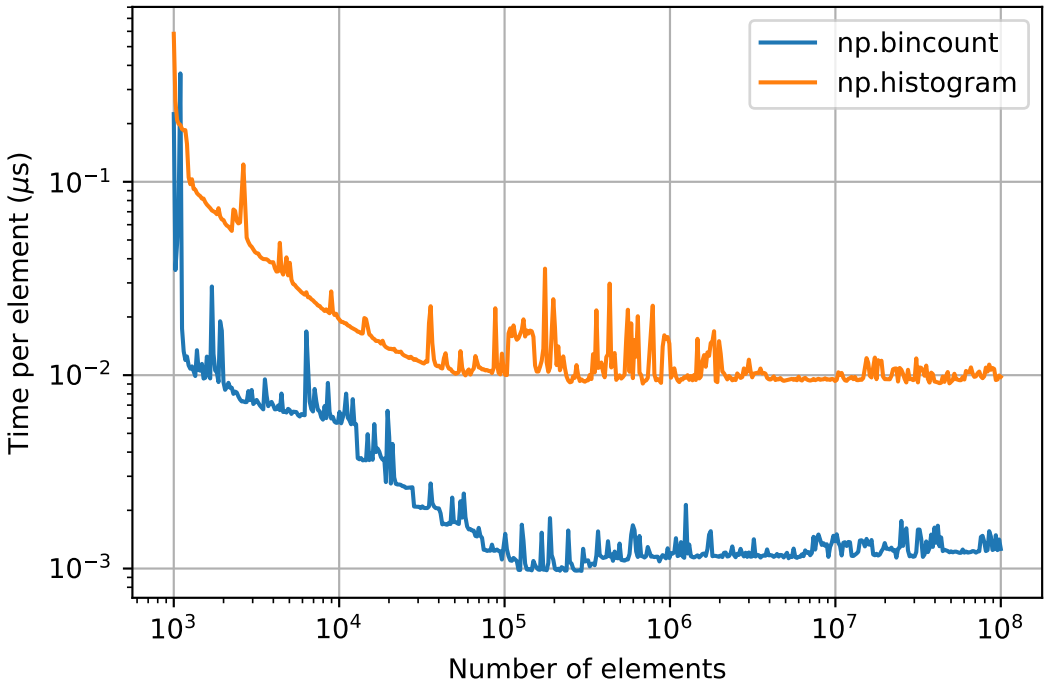

FIGURE 5.2: Comparison of the time per array element versus number of elements in the array for the functions np.histogram and np.bincount. Both cases computed for randomly generated arrays of integers between 0 and 1000, classified in 1000 uniform bins.

The slowest part of this code (in a sense, the bottleneck) is between line 21 and line 29: computing the histograms of cell pair data has by far the largest computational cost. Since this is handled solely by NumPy functions (sorting the cell pairs directly in Python is completely out of the question), the key to optimisation lies in the selection of the right NumPy function and its proper parameters.

The "default" options for this task in NumPy are the functions np.histogram and np.histogram2d (for binning in both $r_{\parallel}$ of $r_{\perp}$). While they allow total flexibility in their binning, they alone take >95% of the computation time. A much faster (and less known alternative) is np.bincount; it is approximately one order of magnitude quicker, as seen in fig. 5.2. However, this speed comes at a cost: it is designed to work only with integers (although a weighting of floats can be applied), and only works in uniform bins of size 1. Therefore, what it does is to return the number of (weighted) occurrences of each integer in a given array. Even if it only works in one dimension, in order to compute a 2d histogram iterating np.bincount over one dimension is still faster than np.histogram2d.

In order to use this function, then, all distances must be scaled to integers; this is done right before the histogram computation, and the scaling factor is a function of both the cell size of the PAUS simulation and the distance binnings for the 2PCFs. Using np.bincount this way, computing the cross-correlation in the PAUS-DESI simulation takes a total of ∼32.6 s: ∼27.3 s spent only in the instances of np.bincount, and ∼5.3 s in the rest of the code. These numbers are considering that all necessary data is already loaded in memory, and for a memory allocation of 5 GB in a single thread of a Intel Xeon E-2144G @ 3.6 GHz.

From these numbers, it can be concluded that further optimisation depends mostly

on the use of a more efficient histogram function. After some research, we found that np.bincount was the best option out of the histogram functions present in common Python packages for data science (e.g., NumPy, Pandas, SciPy), so a significant reduction on computation time probably needs a more specific library. Consequently, there is no need to further optimise the Python part of the code (i.e., algorithm 1).

On the other hand, this also means that parallelisation can only improve reduce the ∼16% of the computing time not spent in np.bincount, since the function can not take any advantage of multithreading; while the current version of the code has been prepared to run in parallel, the gains are fairly small. Another caveat to take into account is that, in order to sort the cell pairs as fast as possible, the np.bincount function needs to "saturate", i.e., to reach the minimum computational time per element in all of its executions. According to fig. 5.2, this limit is reached around just $\sim 10^5$ elements, but still, larger memory allocations imply less times running line 21 to line 29, and reduced net overheads when computing the histograms.

For a more technical approach to the cross-correlation algorithm, the actual core functions of the Python code are displayed and briefly discussed in appendix A.

5.3 Theoretical correlation function

To validate the result of the simulated cross-correlation, as well as to derive the clustering, comparison against a theoretical 2PCF is needed. The first step is to compute the unbiased matter-matter 2PCF from the theoretical matter power spectrum. For this work, this 2PCF has been initially computed as a field depending on two variables, the distances parallel and perpendicular to the line of sight, $r_\parallel$ and $r_\perp$ respectively, with the following expression (e.g. see Hui, Gaztañaga, and Loverde 2007; Gaztañaga, Cabré, and Hui 2009)

$$\xi(r_\parallel, r_\perp) = \frac{1}{2\pi^2}\int_0^\infty dk\, k P_{\rm nl}(k) \frac{\sin\left(k\sqrt{r_\parallel^2 + r_\perp^2}\right)}{\sqrt{r_\parallel^2 + r_\perp^2}} \exp(-k r_{\rm cut}). \tag{5.12}$$

Where $P_{\rm nl}(k)$ is the non-linear matter power spectrum computed with CAMB (Lewis, Challinor, and Lasenby, 2000), and the non-linear modelling of HALOFIT (Peacock et al., 2014). This power spectrum has been computed at the redshift of the hydrodynamic simulation snapshot ($z = 3$), using its cosmology (table 4.1). Regarding other terms, $r_{\rm cut}$ is the radius of the exponential cutoff set in order to avoid large oscillations in the theoretical 2PCF due to small-scale effects that are not represented in its counterpart measured in the simulation. For this study, the chosen value for this cutoff is $r_{\rm cut} = 3$ Mpc/h.

By definition, there is no anisotropy in eq. (5.12), which may make the computation of the 2PCF in two directions seem redundant. Nevertheless, two effects that are to be taken into account will break the isotropy of the function: the smoothing introduced by the binning of the simulated data, and the effect of redshift-space distortions (RSDs).

5.3.1 Smoothing

This effect arises from the fact that correlation is being performed between spatial cells with finite volumes, whose value of the field to cross-correlate is the average over the volume of the cell. If the length of these cells to cross-correlate is equal or smaller than the binning of the correlation estimator (eq. (5.1)), this effect will

be negligible, given that by binning the estimator already averages over a similar length. This is the case for the Lyα forest cells in all directions or the PAUS cells in RA and dec directions, where their length (1.56 Mpc/h, given by the hydrodynamic simulation bins) is smaller than the binning of the estimated cross-correlation (see fig. 6.2).

On the other hand, this effect is not negligible for the redshift direction in PAUS cells, where the mean cell size is 56.25 Mpc/h (since the redshift bins have been merged to simulate PAUS filters). Averaging the Lyα flux in PAUS images over such distances will certainly have an effect on the estimated cross-correlation, which also has to be simulated in the theoretical 2PCF. Considering that the redshift direction in the simulation has a direct correspondence with $r_\parallel$ in eq. (5.12), this smoothing can be emulated by averaging each point in the computed $\xi(r_\parallel, r_\perp)$ field over a length in $r_\parallel$ equal to the average PAUS cell size

$$\bar{\xi}(r_\parallel, r_\perp) = \frac{1}{l_\parallel} \int_{r_\parallel - l_\parallel/2}^{r_\parallel + l_\parallel/2} dr'_\parallel \, \xi(r'_\parallel, r_\perp). \tag{5.13}$$

Where $l_\parallel = 56.25$ Mpc/h. By definition of the 2PCF (eq. (2.3)), $r > 0$, so for $r_\parallel < l_\parallel/2$ this expression changes to

$$\bar{\xi}(r_\parallel, r_\perp) = \frac{1}{l_\parallel} \left[\int_0^{r_\parallel + l_\parallel/2} dr'_\parallel \, \xi(r'_\parallel, r_\perp) + \int_0^{l_\parallel/2 - r_\parallel} dr'_\parallel \, \xi(r'_\parallel, r_\perp) \right]. \tag{5.14}$$

If the 2PCF is interpreted as an average product of cell pairs at a certain distance, such as in the estimator we use, this last expression represents the case where the small Lyα forest cell lies inside the redshift range of the PAUS cell it is being cross-correlated with. The smoothing integral needs to cover the whole $l_\parallel$, but since the distance between cells necessarily has to be non-negative, the integral is truncated in two terms: one for the portion of the PAUS cell at higher redshift than the Lyα forest cell, and another for the portion at lower redshift. Figure 5.3 shows the effect of this 2PCF smoothing (dashed lines) compared to the non-smoothed 2PCF (solid lines) for the three correlation types considered in this work.

5.3.2 Bias and RSDs

In addition to this smoothing effect, bias from the tracers also needs to be taken into account, as well as the effect of RSDs. Since the scales studied in this work are large enough, the only RSD effect considered is the Kaiser effect (Kaiser, 1987).

So far, the unbiased matter-matter 2PCF has been considered (called $\bar{\xi}_{\rm mm}$ henceforth), but the cross-correlation in this work uses Lyα emission and Lyα forest absorption. The power spectrum of a tracer t correlated with itself can be obtained from the unbiased 2PCF with the following expression

$$P_{\rm tt}(k) = b_{\rm tt}^2 (1 + \beta_{\rm tt} \mu_{\vec{k}}^2)^2 P(k). \tag{5.15}$$

Where $b_{\rm tt}$ is the bias of the tracer (Lyα emission and Lyα forest in this case), μ_k is the cosine of the angle between the vector position vector in Fourier space $\vec{k}$ and the line of sight, and $\beta_{\rm tt}$ is the effective RSD parameter. $P_{\rm tt}(k)$ is the power spectrum of the tracer; given that RSDs are included in the expression, it is implied to be in redshift space.

If we are to compute the 2PCF from this $P_{\rm tt}(k)$, we need to take into account the dependency of $\mu_{\vec{k}}$ with $r_\parallel$ and $r_\perp$. For the two-dimensional 2PCF $\mu_{\vec{k}}$ will depend on the values of $r_\parallel$ and $r_\perp$, but for the monopole, parallel and perpendicular correlation,

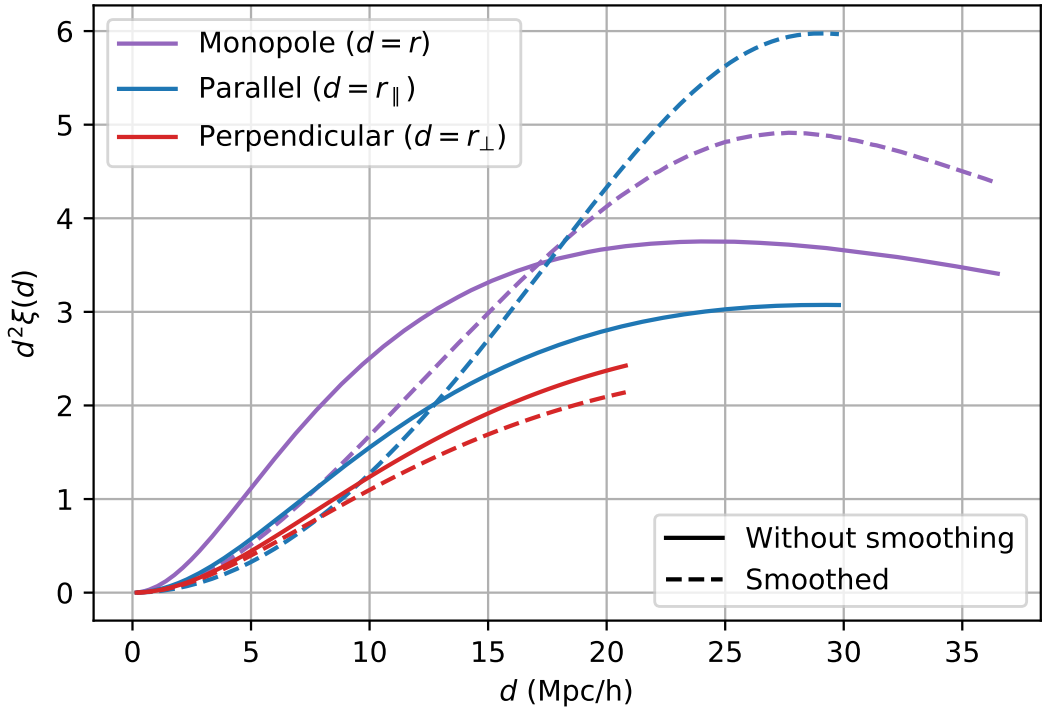

FIGURE 5.3: Theoretical unbiased 2PCFs times squared distance, before and after applying smoothing. The distance range of each one of the 2PCFs is the same as all the results shown in chapter 6. Redshift space distortions have not been considered in this figure.

this dependency vanishes. For the parallel and perpendicular case, $\mu_{\vec{k}} = 1$ and $\mu_{\vec{k}} = 0$ respectively, while for the monopole, after integration in all directions eq. (5.15) becomes

$$\bar{\xi}_{\mathrm{tt}}(r) = b_{\mathrm{tt}}^2 \left(1 + \frac{2}{3}\beta_{\mathrm{tt}} + \frac{1}{5}\beta_{\mathrm{tt}}^2\right)^2 \bar{\xi}_{\mathrm{mm}}(r). \tag{5.16}$$

For this work, two tracers are considered: the Lyα emission (denoted by e) and the Lyα absorption that generates the Lyα forest (denoted by a); the autocorrelation of both tracers will be computed to validate the simulation, as explained in section 6.1. The assumed bias factors b for both tracers are already explained in section 4.2, while the β parameter needs different assumptions for each case.

Given that the Lyα emission has been considered proportional to the square of the matter density field, its β will be the same as the matter density, scaled by the bias factor b_{aa}. Following the approximation in Kaiser, 1987 for linear theory, we find

$$\beta_{\mathrm{ee}} = \frac{\Omega_m(z)^{0.6}}{b_{\mathrm{ee}}}. \tag{5.17}$$

Regarding the Lyα absorption, β_{aa} is independent from b_{aa}, since the nonlinear transformation applied to obtain the Lyα absorption field (eq. (4.1)) does not preserve the flux between real and redshift space. According to Slosar et al., 2011, the value of β_{aa} for different simulations depends on their resolution, with values oscillating between 1 and 1.5 at z$\sim$2.25. Recent observations of BAO with the Lyα forest autocorrelation yield $\beta_{\mathrm{aa}} \sim 1.8$ at $z = 2.34$ (de Sainte Agathe et al., 2019), but also show some evidence of β_{aa} decreasing with redshift. Since the mean redshift of this work ($z \sim 3$) is higher than any of the cited values of β_{aa}, and the hydrodynamic simulation is relatively low-resolution, we decide to adopt the conservative value of $\beta_{\mathrm{aa}} = 1$.

So far, we have discussed how to model the effects of the bias and the RSDs for the autocorrelation of the Lyα emission and the Lyα absorption from the forest; however, the cross-correlation between both tracers also needs to be modeled. In order to do so, we compute the effective biases of each tracer for cross-correlation, $\hat{b}_{\mathrm{t}}$ as the square root of the tracer bias times its RSD factor. For example, for the monopole, following eq. (5.16) the effective bias would be

$$\hat{b}_{\mathrm{t}} = b_{\mathrm{tt}} \left(1 + \frac{2}{3}\beta_{\mathrm{tt}} + \frac{1}{5}\beta_{\mathrm{tt}}^2\right). \tag{5.18}$$

Therefore, the monopole cross-correlation can be expressed as

$$\hat{\bar{\xi}}_{\mathrm{ea}}(r) \simeq -\hat{b}_a(r)\hat{b}_e(r)\bar{\xi}_{\mathrm{mm}}(r). \tag{5.19}$$

Where the minus sign comes from the fact that the cross-correlation is between an emission and an absorption field. Table 5.1 sums up the b_{tt} and β_{tt} considered for both Lyα emission and absorption, as well as the resulting effective bias $\hat{b}_{\mathrm{t}}$ for the monopole.

5.3.3 Monopole, parallel and perpendicular 2PCF

This two-dimensional 2PCF computed with eq. (5.12) has to be converted to a 2PCF depending solely on a single distance parameter, either the total distance between cell pairs $r = (r_\parallel^2 + r_\perp^2)^2$, or the parallel/perpendicular distances, in order to be

TABLE 5.1: Bias factor b_{tt} and RSD parameter β_{tt} considered for Lyα emission and absorption, as well as the resulting effective bias for cross-correlation $\hat{b}_t$, according to Eq. 5.18.

	b_{tt}	β_{tt}	$\hat{b}_t$ monopole
Lyα emission	2.000	0.488	2.343
Lyα absorption	0.336	1.000	0.557

compared to the estimator defined in eq. (5.1). The estimator could also be defined as a function of both $r_\parallel$ and $r_\perp$, however, this would greatly reduce the number of cell pairs available per bin, and thus the SNR of the measured cross-correlation.

For the monopole 2PCF, this has been performed by computing $\bar{\xi}_{mm}(r_\parallel, r_\perp)$ in a very fine uniform grid of $r_\parallel$, $r_\perp$ values, and then averaging these values in bins of total distance $r = (r_\parallel^2 + r_\perp^2)^2$. Regarding the parallel and perpendicular 2PCFs, they have been obtained from the theoretical two-dimensional 2PCF simply by numerical integration, according to the following expressions

$$\begin{aligned} \bar{\xi}_{mm}(r_\parallel) &= \frac{1}{R_\perp} \int_0^{R_\perp} dr_\perp \, \xi_{mm}(r_\parallel, r_\perp) \\ \bar{\xi}_{mm}(r_\perp) &= \frac{1}{R_\parallel} \int_0^{R_\parallel} dr_\parallel \, \xi_{mm}(r_\parallel, r_\perp), \end{aligned} \tag{5.20}$$

where $R_\parallel$ and $R_\perp$ are the maximum binning distances used by the estimator in eq. (5.1) for the parallel and perpendicular directions. These 2PCFs, unlike the monopole, depend on the total range over which the correlation is computed, which makes them less suitable for comparison of the results against the theory. Consequently, only the monopole 2PCF will be used to compare the results of the simulation against the theory in chapter 6.

Chapter 6

Results and conclusions

The results presented in this thesis are divided in three parts. First, in section 6.1 we compare the results of our Lyα IM simulation, without noise nor foregrounds, against the theoretical 2PCF computed in section 5.3. This allows us to validate our model, as well as verify the actual bias of the tracers in our simulation (with RSDs included, as defined in eq. (5.18)). Besides, it can also be interpreted as an ideal (and thus, unfeasible) result where foregrounds are completely removed and noise is negligible.

Second, in section 6.2 the probability of a detection (SNR>3) when adding foregrounds and instrumental noise using the uncorrelated extrapolation (table 4.5) is explored at different scales. Four cases of cross-correlation are explored: PAUS-eBOSS, PAUS-DESI, and the cross-correlation of DESI with two hypothetical extensions of PAUS (an increase in exposure time, PAUS deep, and in survey field, PAUS extended). These results can be considered and optimistic estimate, assuming that further work has been dedicated to understand and mitigate the noise correlations that appear in PAUS images.

Third, in section 6.3 we summarise the same results, but using the correlated noise directly measured from PAUS images, instead of the uncorrelated extrapolation; this is a realistic estimate with our current understanding of noise correlation and the reduced images already available. Finally, with conclude in section 6.4 summarising the methodology and results of the thesis, as well as laying out some possible future work that can be developed based on the results obtained here.

6.1 Cross-correlation without noise or foregrounds. Comparison against theory

In order to compare the cross-correlation results against the theoretical prediction (and thus validate that the cross-correlation results are sound), the actual effective biases of the tracers of the hydrodynamic simulation need to be measured and compared against the expected values from table 5.1.

These effective biases of the simulation have been determined by correlating the emission/absorption arrays of the hydrodynamic simulation (fig. 4.3) with themselves, using the same binning as in the PAUS-eBOSS/DESI simulations (wide redshift bins for PAUS, only Lyα forest cells for eBOSS/DESI). No foregrounds or noise were added for this correlation, since they do not have the same physical units, and the purpose of this calculation is just to determine the real bias while testing that the binning of the simulation and the smoothing effect are properly taken into account. Considering eq. (5.16) and eq. (5.18), the effective bias of the tracer can be estimated from the smoothed theoretical prediction $\bar{\xi}_{\mathrm{mm}}$ and the estimated correlation of the tracer $\hat{\bar{\xi}}_{\mathrm{tt}}$ with

$$\hat{b}_t(r) = \sqrt{\frac{\hat{\xi}_{\mathrm{tt}}(r)}{\bar{\xi}_{\mathrm{mm}}(r)}}. \tag{6.1}$$

Where t is any tracer, and the expression has been considered only for the monopole 2PCF. The results of this bias determination are displayed in fig. 6.1. The error of the bias at all distance bins is simply the propagated error of the cross-correlation; any error that could be included in the theoretical 2PCF (e.g., cosmic variance) has been considered negligible.

As can be seen in fig. 6.1, the actual measured $\hat{b}_{\mathrm{a}}$, around 0.4, is smaller by $\sim$25% than the expected value from table 5.1, but for the emission field $\hat{b}_{\mathrm{e}}$ is actually really close to the predicted value. This is to be expected: the Lyα emission field is proportional to the square of the matter density field, which already gives an exact value of the bias, and the effect of RSDs in this case are well described by linear theory. However, for the Lyα forest both b and β are uncertain (especially the latter), and the reference values available come from measurements/simulations at lower redshifts, so such discrepancies are reasonable.

With these measured biases, the simulated cross-correlation can be compared to the theoretical 2PCF by applying eq. (5.19), but instead of using the effective bias values of table 5.1, we apply the measured effective biases from fig. 6.1 to each distance bin.

A comparison of the simulated cross-correlation, without either foregrounds or instrumental noise, to the theoretical 2PCF with the measured biases is displayed in fig. 6.2. Only the monopole 2PCF is shown, since the parallel and perpendicular 2PCF depend on the range in which the 2PCF is computed, as shown in eq. (5.20). No foregrounds or instrumental noise have been added both to ensure a good SNR to validate our model, and because the noise bias described in eq. (5.11) would also need to be corrected to compare the simulation against theory.

With no foregrounds or instrumental noise, there is a clear detection of cross-correlation at $r > 30$ Mpc/h with DESI, and several bins show a clear detection up to $r \sim 30$ Mpc/h with eBOSS. For all the bins with a detection, the errorbars of the theoretical prediction and the actual cross-correlation overlap; this validates the simulated cross-correlation. Besides, this also proves that, for an ideal case without any other sources of noise, this cross-correlation could be used to constrain either the bias of the tracers or the 2PCF on scales up to $\sim$30 Mpc/h.

Nevertheless, when the foregrounds and the instrumental noise from PAUS are added to the simulation, the general SNR of the cross-correlation drops greatly. Therefore, instead of simulating the cross-correlation and comparing to the theory (assuming that a detection is almost certain), a different approach has been taken to evaluate the probability of a detection.

6.2 Cross-correlation with uncorrelated noise: Probability of detection

6.2.1 PAUS-eBOSS/DESI

As explained in chapter 4, a simulation of the cross-correlation contains three stochastic elements: the instrumental noise in PAUS images, the Gaussian noise inserted in the Lyα forest simulation, and the quasar cells in eBOSS/DESI that determine the Lyα forest cells to be sampled (following the redshift distributions in fig. 3.7). Nevertheless, the noise in the Lyα forest is clearly subdominant (its addition does not

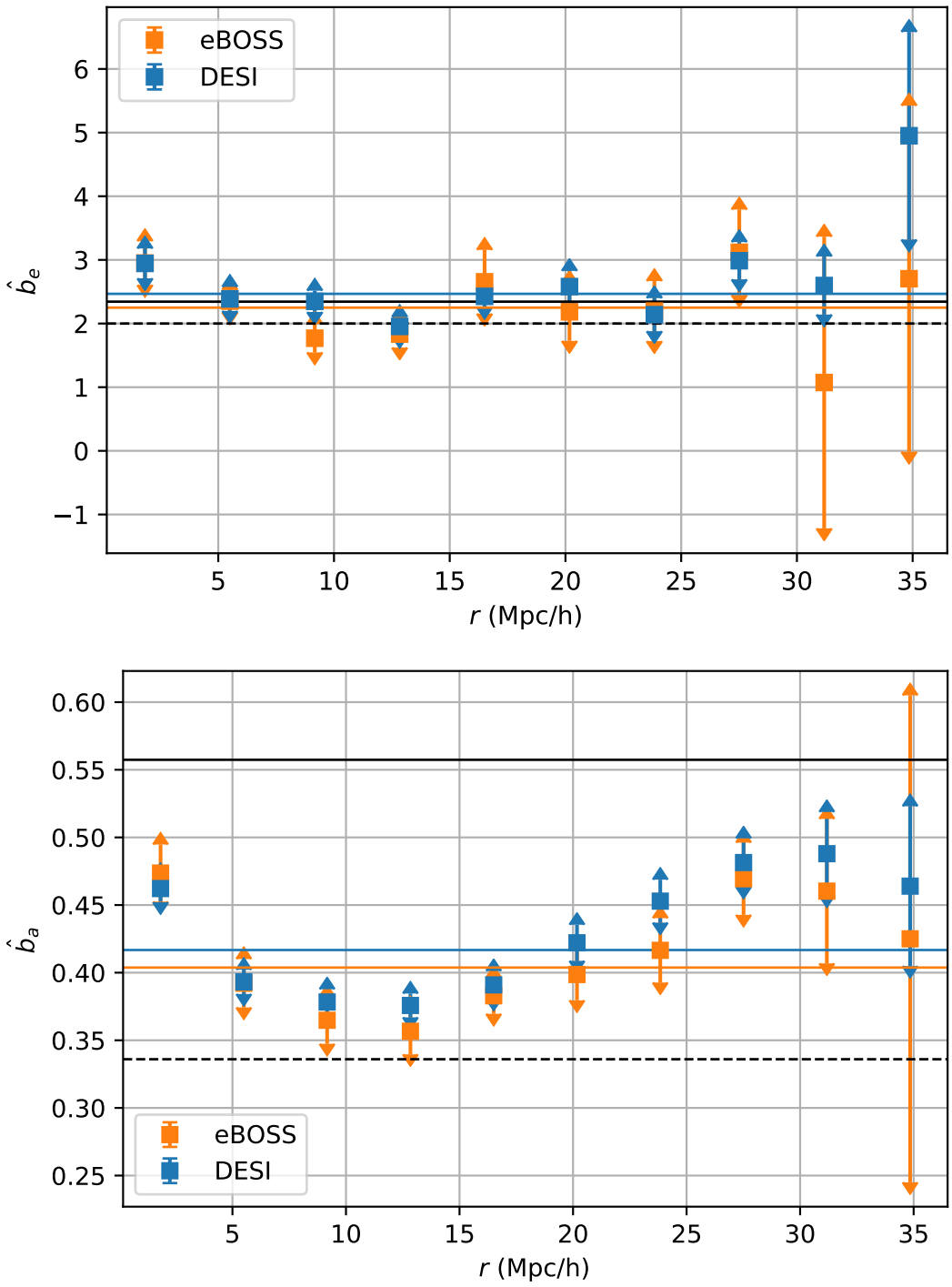

FIGURE 6.1: Emission bias (top panel) and absorption bias (bottom panel), measured as described in eq. (6.1), for correlations using eBOSS and DESI Lyα forest binnings. Solid black line represents the approximate effective bias from section 4.2.1; dashed black line represent the theoretical bias without considering RSDs ($\beta = 0$). Coloured horizontal lines are the weighted average of the recovered effective bias for the respective surveys.

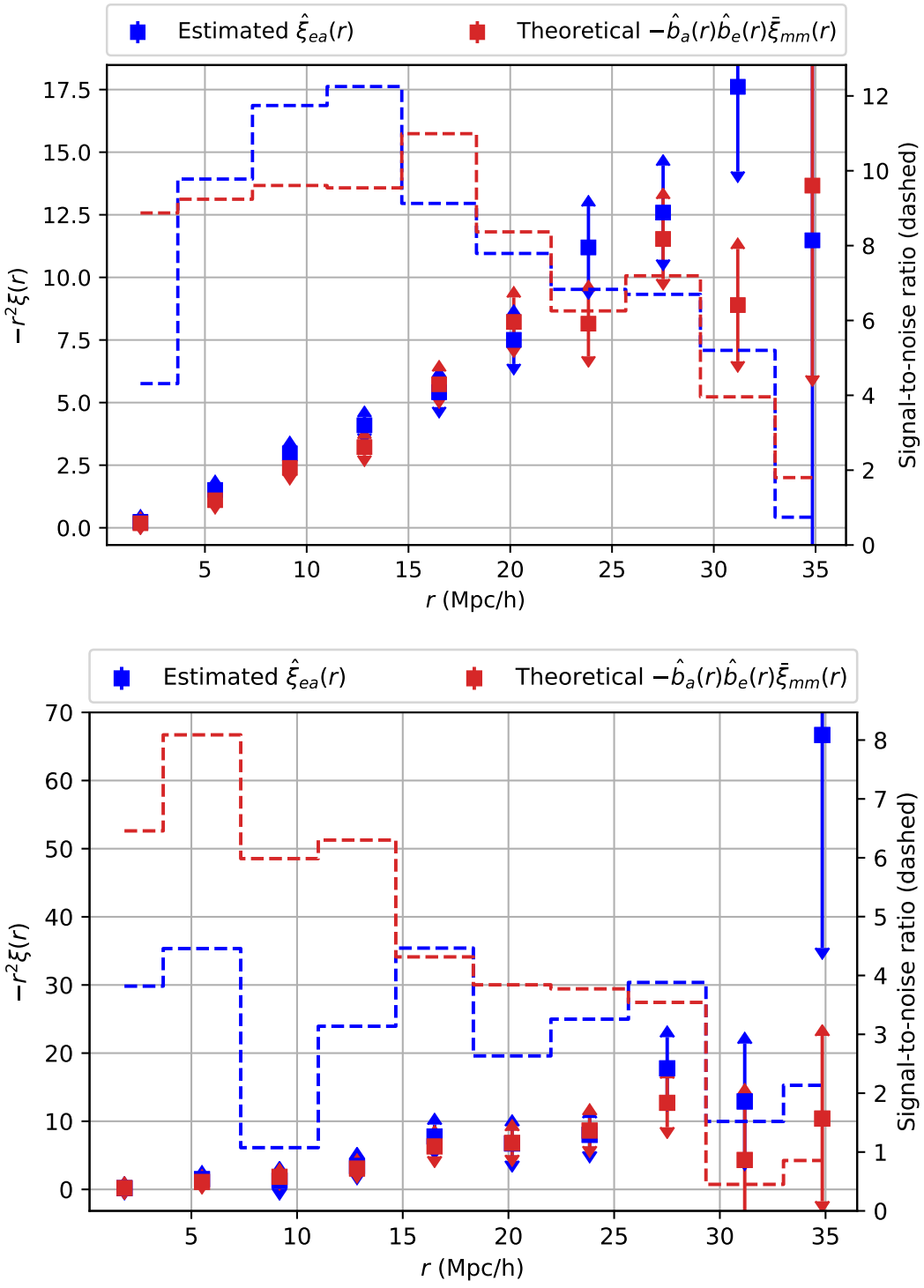

FIGURE 6.2: Comparison of the simulated cross-correlation with foregrounds, without instrumental noise, and the theoretical 2PCF with the measured biases. The points with errorbars represent the cross-correlation values (left y axis), while the dashed line represents the SNR of each distance bin (right y axis). *Top panel:* PAUS-DESI. *Bottom panel:* PAUS-eBOSS.

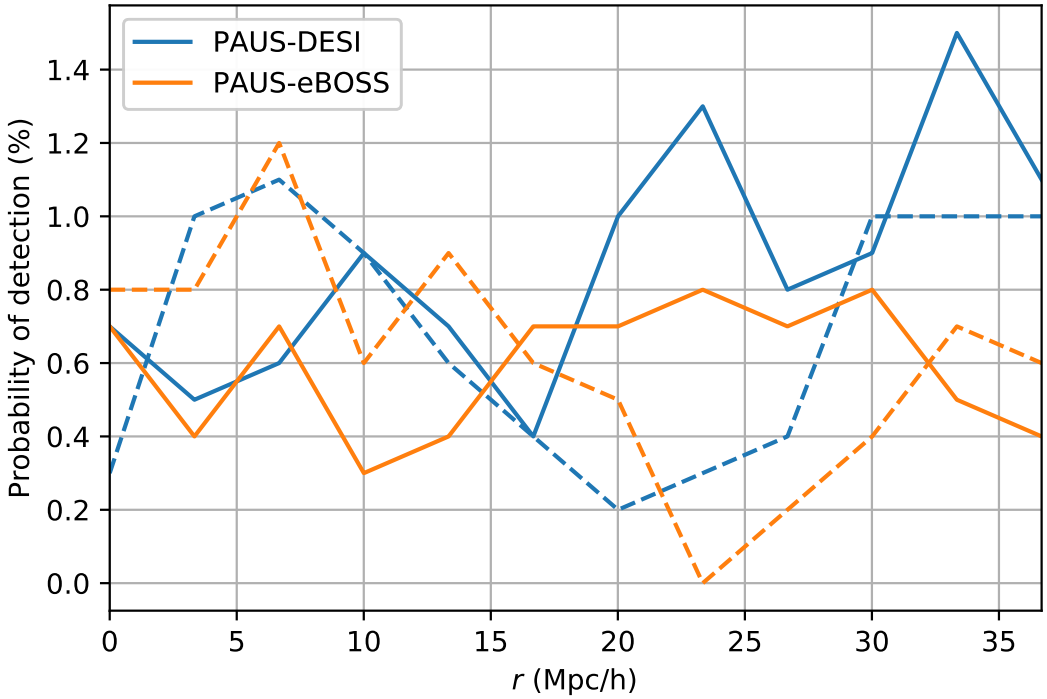

FIGURE 6.3: Probability of a detection for the monopole 2PCF as a function of distance in the simulated cross-correlations PAUS-eBOSS and PAUS-DESI, for 1000 different realisations of instrumental noise+Lyα forest. Solid line displays the actual probability of any detection, dashed line shows the probability of a spurious detection.

alter the results), so we will discuss only the PAUS noise and the quasar positions from now on.

Without the instrumental noise, different realisations of the Lyα forest quasar positions do not modify significantly the cross-correlation results. Nevertheless, when the instrumental noise (using the uncorrelated extrapolation) is added to the PAUS simulation, the SNR of the cross-correlation heavily decreases, up to the point of a detection (SNR>3) depending on the realisation of the noise and the Lyα forest (i.e., the SNR is not consistent between different runs of the simulation pipeline). Fixing one of these stochastic elements (either the Lyα forest position or the instrumental noise) does not gives consistent results either.

Therefore, the approach we have taken is to simulate the cross-correlation 1000 times, with different realisations of the instrumental noise and the Lyα forest quasars each time, and compute the probability of detection as the fraction of the realisations where at least one bin has SNR>3. For each one of the realisations, the monopole, parallel and perpendicular 2PCF have been computed using 12 uniform distance bins; finer distance bins would result in empty bins (without any cell pairs) for some cases.

In addition to this, the cross-correlation has also been computed another 1000 times for each case, but with the Lyα emission in PAUS images mirrored both in the RA and dec axes. This way, the actual cross-correlation between the simulated PAUS images and the Lyα forest should be null, as one of the two signals has been inverted, so any detection that results from this cross-correlation is inherently spurious. In

fact, the 2PCF has only been computed with 12 uniform bins partially because applying other uniform binnings seemed to increase the spurious detections without a larger net increase in real ones.

Figure 6.3 displays the probability a detection in the monopole 2PCF for these 1000 runs for PAUS-eBOSS and PAUS-DESI: the solid line represents the probability of any detection from the real cross-correlation (this includes the spurious ones), and the dashed line the probability of a spurious detection (for the cross-correlation with inverted Lyα emission). Figure 6.4 shows analogous information, but for the parallel (top panel) and perpendicular (bottom panel) 2PCF.

As it would be expected, cross-correlation with DESI noticeably increases the detection probability for all three 2PCFs; however, the probability of any detection is still very small. In the PAUS-eBOSS case, the solid line is barely above the dashed one, which implies the the probability of a detection is almost negligible. In PAUS-DESI, there are some regions where the probability of any detection is clearly above the dashed line; these are the cases that we will discuss.

First, the monopole/parallel 2PCF and the perpendicular 2PCF seem to better sample different scales: the parallel 2PCF has a lower probability of detection at small scales, and shows an increase in detection probability around 10 Mpc/h, while the perpendicular 2PCF has exactly the opposite behaviour. The cause of these contrasting trends in the detection probability for different 2PCFs is the smoothing effect that PAUS filters have in the parallel (redshift) direction, displayed in fig. 5.3.

This figure shows the 2PCF multiplied times its respective distance squared; given that our cross-correlation algorithm selects a cube of surrounding PAUS cells to compute the respective pairs of each forest cell, the surface of this cube (and thus the amount of available pairs for each distance bin) will grow as r^2. Therefore, $r^2\xi(r)$ can be interpreted as the 2PCF weighted by the number of pairs sampled, and thus is a better predictor of SNR than $\xi(r)$ alone.

In the parallel cross-correlation, and to a lesser extent, the monopole 2PCF, smoothing decreases the absolute value of the 2PCF at scales of 10 Mpc/h and 15 Mpc/h respectively, while at larger scales the 2PCFs are increased (at least, as far as the size of the hydrodynamic simulation allows to compute the 2PCF, 30-35 Mpc/h). This trend matches almost perfectly the detection probabilities in fig. 6.3 and fig. 6.4, with sharp increases in the monopole and parallel 2PCF at the same scales.

On the other hand, the perpendicular 2PCF in fig. 5.3 shows a smaller decrease, even when going to larger scales than the ones depicted in fig. 5.3; this small effect of the smoothing results in higher detection probabilities at smaller scales, where the 2PCF has higher absolute values. This result shows that the parallel and perpendicular 2PCFs are highly complementary, and both should be taken into account for any future observational studies of Lyα IM with PAUS (or similar surveys) in order to maximise the probability of a detection at all scales.

Nevertheless, it is worth noting that, even in these regions where the detection probability increases, the difference with the spurious detection probability is really modest, and outside of these regimes the probability of an spurious detection is larger. While this should not be technically possible, it is due to the fact that these two probabilities come from two different finite sets of realisations, and thus they have an intrinsic variance. If anything, it can be interpreted as the effective probability of a non-spurious detection being null.

Furthermore, the total probability of any detection (regardless of the kind of 2PCF and the binning) has also been computed, considering that any realisation where one or more bins in any 2PCF had SNR>3 was an effective detection. These results are summarised in the two upper rows of table 6.1; the probability of a real

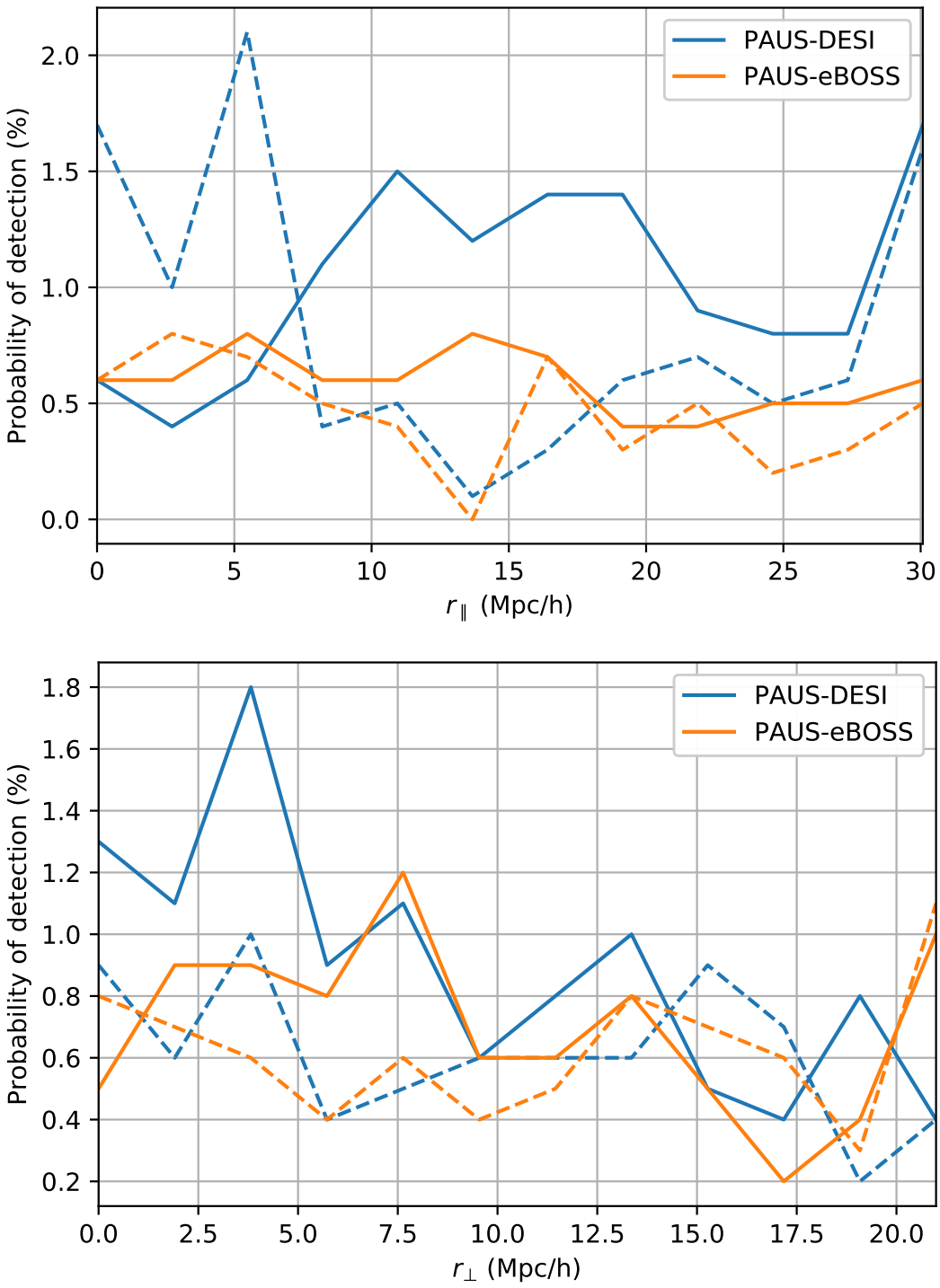

FIGURE 6.4: Probability of a detection as a function of distance in the simulated cross-correlation PAUS-eBOSS and PAUS-DESI, for 1000 different realisations of instrumental noise+Lyα forest. Solid line displays the actual probability of any detection, dashed line shows the probability of a spurious detection. *Top panel:* Parallel 2PCF. *Bottom panel:* Perpendicular 2PCF.

TABLE 6.1: Probability of any detection for simulated cross-correlations PAUS-eBOSS and PAUS-DESI, using the uncorrelated PAUS noise extrapolation.

Surveys	Any detection	Spurious	Real
PAUS-eBOSS	18.7%	16.9%	1.8%
PAUS-DESI	24.7%	20.2%	4.5%
PAUS deep-DESI	39.5%	24.2%	15.3%
PAUS extended-DESI	36.8%	27.8%	9.0%

detection is simply the difference between the probability of any detection (percentage of the 1000 realisations that yielded a detection) an the probability of a spurious detection (percentage of the 1000 realisations with inverted Lyα emission where a detection happened). With this approach, we assume that the probability of a real detection of Lyα cross-correlation and a spurious one are independent processes.

When considering these results, it is important to take into account that in this study no weightings to improve SNR of the estimator in eq. (5.1) have been considered, and only a uniform binning has been applied for the 2PCFs computation (which are common methodologies to optimise the SNR in an observational analysis). Nevertheless, the detection probabilities are still very small, with the probability of a spurious detection being far higher than an actual one in both cases; it is safe to assume that any statistical approach to increase SNR is unlikely to yield significantly better results.

6.2.2 Hypothetical cases: PAUS deep, PAUS extended

In addition to the PAUS-eBOSS and PAUS-DESI simulations, two hypothetical cases have also been considered: PAUS deep, a survey with the same field coverage, but complete up to a magnitude deeper ($i_{\rm AB} < 24$), and PAUS extended, with the same exposure time as current PAUS, but a larger angular area of 225 deg^2. These hypothetical PAUS cases have only been cross-correlated with the DESI simulation, since eBOSS would be rendered obsolete by DESI before such hypothetical surveys could be finished.

PAUS deep has been simulated analogously to PAUS, with the sole difference being the instrumental noise, now reduced by a factor of $\sqrt{6}$, as displayed in table 4.4 ($\sigma_{\rm 18exp\,abs}$). Regarding PAUS extended, the Lyα emission array has been repeated in a mosaic of 3x3, instead of 2x2, thus yielding an angular coverage of $\sim$225 deg^2. To cover this mosaic of Lyα emission, the foregrounds array has been repeated and rotated for the first 4 iterations; after that, it has been mirrored in RA direction and repeated until the 3x3 mosaic has been filled. This gives 8 possible combinations of Lyα emission-foregrounds: the 4 rotations of the foreground array plus the 4 mirrored rotations, which sets a limit on the maximum area we can simulate in this study. In fact, the 3x3 mosaic already has one redundant combination of Lyα emission+foregrounds (since it is composed of 9 realisations). Simulating even larger areas would result in largely redundant foregrounds, which would provide too optimistic results given that the same combination of Lyα emission+foregrounds would be sampled multiple times.

The probability of detection for 1000 realisations of these simulations, together with original PAUS-DESI simulation, is shown for the monopole 2PCF in fig. 6.5, and

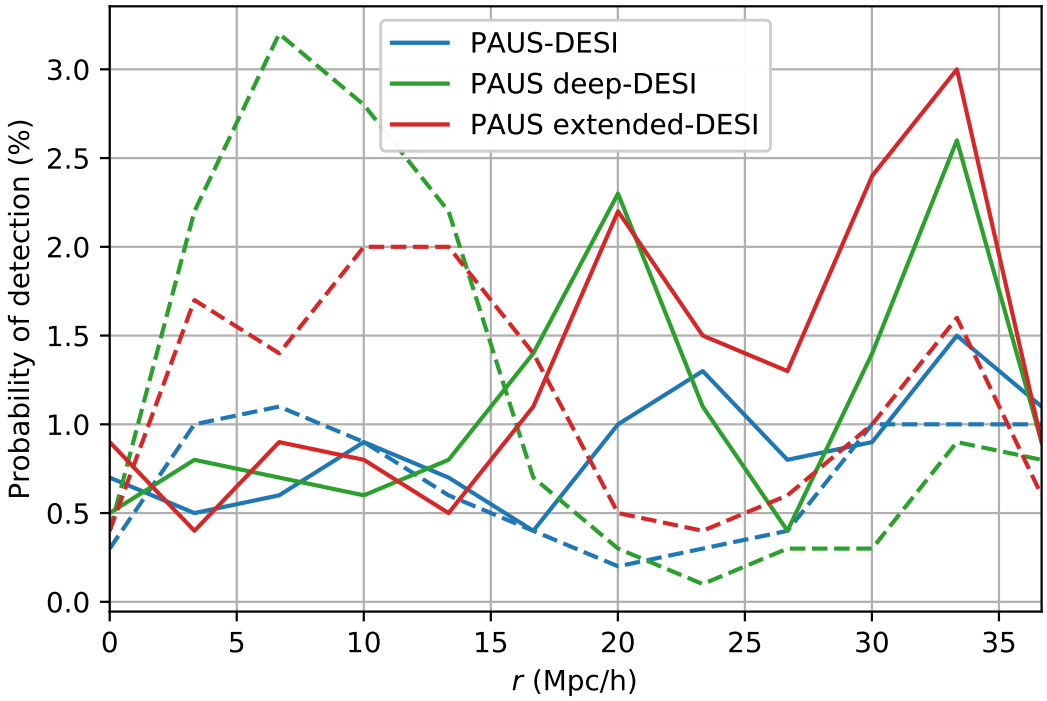

FIGURE 6.5: Probability of a detection for the monopole 2PCF as a function of distance in the simulated cross-correlation of hypothetical extensions of PAUS and DESI, for 1000 different realisations of instrumental noise+Lyα forest. PAUS deep refers to a survey complete up to $i_{\rm AB} < 24$ (exposure time x6), while PAUS extended refers to a total survey area of 225 deg^2. Solid line displays the actual probability of any detection, dashed line shows the probability of a spurious detection.

its parallel and perpendicular counterparts in fig. 6.6. Table 6.1 displays the probability of any detection (two lower columns); the probability of spurious detections has also been computed following the same methodology.

Both cases show a noticeable increase in the probability of any detection, which is now close to 40%. However, the probability of spurious detections also increases, which diminishes the net gain in the probability of any detection. Overall, spurious detections are more likely, but the probability of a real detection is only smaller by a factor of few (less than 2 for PAUS deep-DESI, and approximately 3 for PAUS extended-DESI).

The same complementary trend is observed in fig. 6.5 and fig. 6.6, with the perpendicular 2PCF sampling better at scales below 10 Mpc/h, while the monopole and parallel 2PCF have a much higher chance of detection at larger scales. For these last two 2PCFs, PAUS extended seems to provide a much higher increase of probability of detection (an increase by a factor of 2-3) at distances larger than 10 Mpc/h, while the improvement of PAUS deep compared to original PAUS is much smaller. The perpendicular 2PCF at small scales, however, shows similar improvement with either PAUS deep or PAUS extended. PAUS deep, however, seems to perform much better at small scales with the perpendicular 2PCF.

PAUS deep would need 6 times the observation time from current PAUS to observe the same area (going from 3 exposures for each pointing to 18), while PAUS extended only would need 2.25 times the observation time to be carried out (since 225 $\deg^2$ are being observed instead of 100 $\deg^2$, with the same exposure time per pointing). Since PAUS deep seems to provide better detection probabilities (by a factor of $\sim$1.5), but also requires almost twice the observational time, it is difficult to assess which strategy is more time-efficient for a Lyα IM detection. Nevertheless, it seems clear that increasing exposure time yields better results on small scales, and observing a larger field increases the detection probability at larger scales.

6.3 Cross-correlation with correlated noise: Probability of detection

So far in this section, we have only discussed the results for the optimistic case where it is assumed that methods are developed to remove the noise correlation the photometric imaging of PAUS. However, this is not the current case, and while a noticeable reduction in the noise correlation might be achieved (since this is an active area of research in other IM applications, such as 21 cm IM, e.g. Liu et al., 2009), using the correlated noise is the most realistic approach for now.

Table 6.2 shows the probability detection results, following the same methodology as in section 6.2, but applying the correlated noise to the simulation of PAUS images. The probability of a detection greatly decreases in both cases, to the point of being negligible in all cases but PAUS deep-DESI, where it is close to 2%. In fact, for the PAUS-eBOSS case we actually obtain more spurious detections with the inverted Lyα signal that total detections with the proper cross-correlations; a clear sign of all of them being spurious.

6.4 Conclusions and future work

In this thesis the possibility of performing Lyα IM by cross-correlation of spectroscopic Lyα forest data with the background of narrow-band images from PAUS has

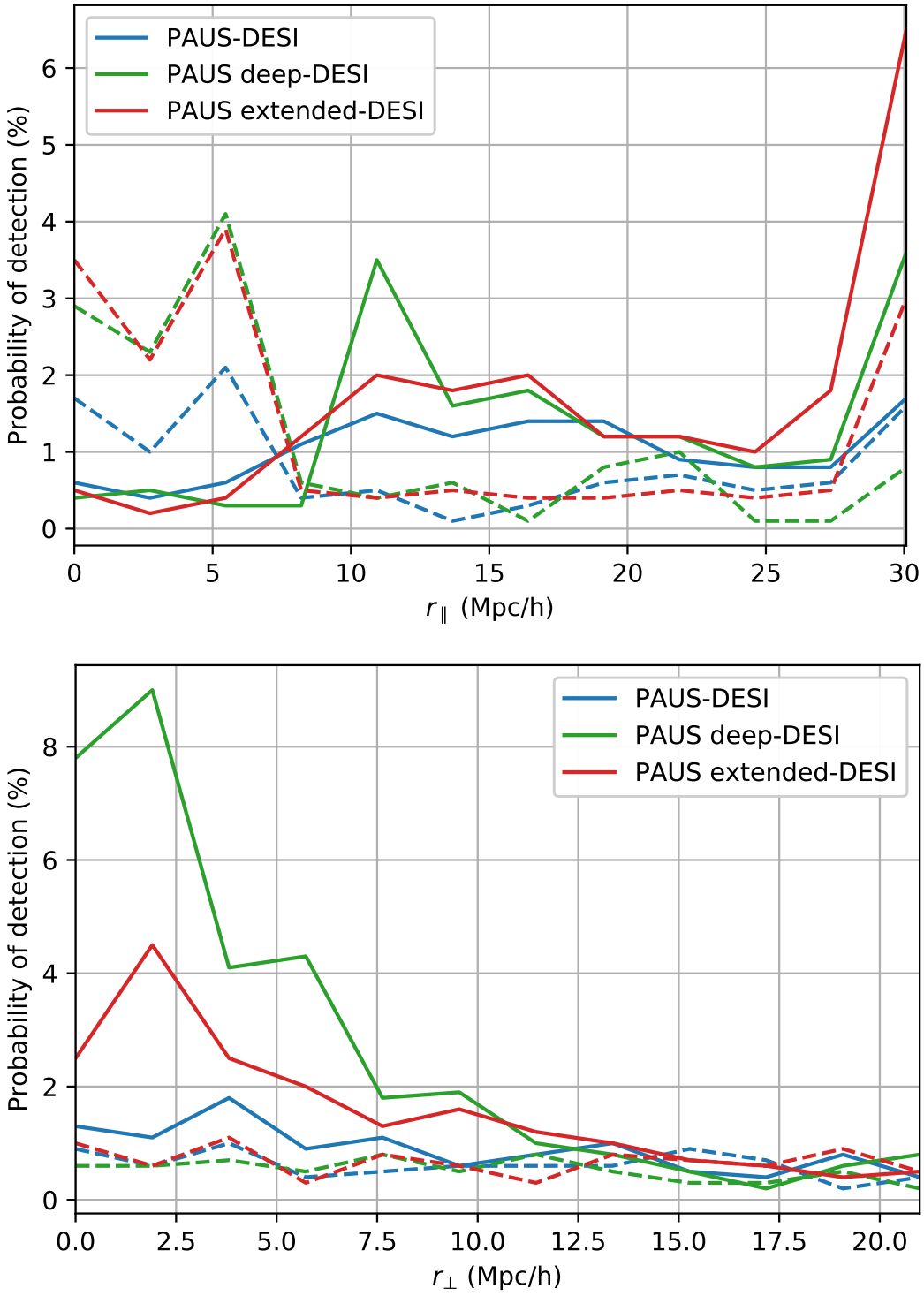

FIGURE 6.6: Probability of a detection as a function of distance in the simulated cross-correlation of hypothetical extensions of PAUS and DESI, for 1000 different realisations of instrumental noise+Lyα forest. PAUS deep refers to a survey complete up to $i_{\rm AB} < 24$ (exposure time x6), while PAUS extended refers to a total survey area of 225 deg^2. Solid line displays the actual probability of any detection, dashed line shows the probability of a spurious detection. *Top panel:* Parallel 2PCF. *Bottom panel:* Perpendicular 2PCF.

TABLE 6.2: Probability of any detection for all the considered cases, using the actual correlated PAUS noise.

Surveys	Any detection	Spurious	Real
PAUS-eBOSS	15.2%	17.5%	0.0%
PAUS-DESI	18.0%	17.9%	0.1%
PAUS deep-DESI	18.5%	16.8%	1.7%
PAUS extended-DESI	18.6%	18.4%	0.2%

been simulated and evaluated. Lyα forest emission and absorption have been simulated from a hydrodynamic simulation of size 400 Mpc/h designed for the study of the IGM (Cisewski et al., 2014; Ozbek, Croft, and Khandai, 2016; Croft et al., 2018). The foregrounds in PAUS images have been simulated from a lightcone mock catalogue made from the Millennium Simulation with the WMAP7 cosmology (Guo et al., 2013), and using the GALFORM (Gonzalez-perez et al. 2014) semi-analytical model. SED templates (Blanton and Roweis, 2006) have been fitted using non-negative least squares to the broad-band data of this mock catalogue in order to achieve PAUS spectral resolution for these foregrounds.

Instrumental noise has been considered for both the simulated PAUS images (measured directly from reduced and stacked PAUS science images) and the Lyα forest spectroscopic data (extracted from Chabanier et al., 2019). Two different cases for the PAUS instrumental noise have been simulated: one with an optimistic uncorrelated noise extrapolation (assuming that noise correlation is mitigated in future work), and another with the current noise levels directly measured, which show a clear correlation at the pixel size of our simulation.

The theoretical 2PCFs (monopole, parallel and perpendicular correlations) have been computed with the derivation shown in Gaztañaga, Cabré, and Hui (2009) from the matter power spectrum, obtained using CAMB (Lewis, Challinor, and Lasenby, 2000). The smoothing of these theoretical 2PCFs due to the large redshift bins for Lyα emission in PAUS narrow-band images has been simulated, and the biases and RSDs of both Lyα emission and absorption have been measured by comparing the theoretical monopole 2PCF to the correlation of the Lyα absorption and emission arrays, using the same spatial binning as the PAUS-DESI cross-correlation. Furthermore, a cross-correlation code in Python has been developed from scratch and optimised to compute the cross-correlations between these simulated datasets; with proper formatting of observational data it could also be run on both actual observed PAUS images and Lyα forest spectra.

Regarding the results, the simulated cross-correlations without foregrounds or instrumental noise show that, despite the redshift smoothing of Lyα emission in PAUS images, and the limited fraction of space sampled by Lyα forest data, the theoretical monopole 2PCF can be recovered, and the bias of both Lyα emission and absorption can be measured. This shows the validity of this technique in an ideal case to both place constraints on the 2PCF and the bias of the extended Lyα emission or the Lyα forest.

Nevertheless, a bias has been identified in the cross-correlation estimator when cross-correlating fields with noise with mean larger than zero (such as the foregrounds and the instrumental noise for this case). This noise bias, while not affecting the SNR, should be taken into account if constraints such as the Lyα emission bias or the Lyα mean luminosity are to be derived from cross-correlation. A constrained model of the foregrounds and other noise sources average values would be needed;

conversely, assuming a known bias and expected Lyα luminosity this same cross-correlation could be used to place constraints on foregrounds emission.

When the cross-correlation is run with the instrumental noise and foregrounds in PAUS images, SNR greatly decreases, up to the point where not all realisations yield a detection. A realisation of this cross-correlation contains three stochastic elements: the instrumental noise of PAUS and the Lyα forest, derived from a random Gaussian distribution (although negligible in the second case), and the positions of the quasars, drawn from the quasar redshift distribution of eBOSS/DESI. Fixing one of these stochastic elements does not provide consistent SNR either, so the probability of a detection (i.e., the cross-correlation reaching a certain SNR threshold) has been evaluated using a purely frequentist approach.

In order to determine this probability, 1000 realisations of the simulated cross-correlations have been carried out with different realisations of both instrumental noise and quasar positions, and for each one the monopole, parallel and perpendicular 2PCFs have been computed for 12 uniform distance bins. Moreover, another 1000 realisations have been computed with mirrored Lyα emission in PAUS images, to determine the probability of spurious detections.

Considering a detection threshold of SNR>3, and under the uncorrelated PAUS noise assumption, Lyα emission has been detected in only 1.8% of PAUS-eBOSS simulations and 4.5% of PAUS-DESI simulations. These percentages increase to 15.3% and 9.0% with two hypothetical PAUS extensions: PAUS deep (going up to $i_{\rm AB} < 24$ instead of $i_{\rm AB} < 23$), and PAUS extended (observing 225 deg^2 instead of 100 deg^2). Nevertheless, in all cases the probability of a spurious detection is higher, and when including the correlated PAUS noise instead, the higher probability of a real detection (PAUS deep-DESI), is just 1.7%. These results clearly show that, even if noise correlation was to be mitigated and PAUS observation time extended, the cross-correlation of the images background with Lyα forest data is unlikely to yield a detection, and if such a detection happens the most likely scenario is that it is spurious.

Despite these negative results, some valuable conclusions can still be extracted. First, the perpendicular and parallel 2PCFs show complementary behaviours: the former has relatively high detection probabilities at scales up to 10 Mpc/h, while the latter displays a non-negligible probability of detection at scales larger than 10 Mpc/h. These different trends are due to the smoothing of the 2PCF in redshift direction, which affects far more the parallel 2PCF than its perpendicular counterpart. Second, this smoothing effect has been properly modelled and recovered when comparing the noiseless correlation to theory, so larger smoothing lengths can be accounted for, and the scales where they will maximize the detection probability, and thus the SNR, can also be predicted.

These two results point out to the fact that broad-band photometric surveys, with angular coverages one or even two orders of magnitude larger than PAUS, such as DES (Abbott et al., 2018) or SDSS (Ahumada et al., 2020) may be more promising for Lyα IM. This is because their main drawback compared to narrow-band surveys (redshift smoothing) has been properly modelled and reproduced, and increasing survey area has been shown to be an effective strategy to increment the detection probability. In fact, since spectroscopic surveys rely on broad-band data for their target selection, the entirety of eBOSS and DESI has already been mapped by broad-band surveys; therefore, all of their footprints could be used for Lyα IM.

Consequently, a possible (and more promising) follow-up to this work would be the simulation Lyα IM with the methodology presented here, but using broad-band data instead of PAUS. The main challenge to overcome in this case, would be the simulation of a much larger angular area, and thus simulated volume (e.g,

the PAUS survey footprint is approximately $\sim$100 deg^2, while the DESI coverage is 14,000 deg^2).

A hydrodynamic simulation (necessary to model Lyα emission and absorption) two orders of magnitude larger than the one used in this thesis simply does not exist; in fact, one of the largest state-of-the-art simulations, IllustrisTNG (Springel et al., 2018) has a maximum diameter of 300 Mpc, approximately half of the size of our simulation (400 Mpc/h, which covers just $\sim$1/4 of PAUS area). Fortunately, recent applications of machine learning in cosmological simulations allow to recreate the HI distribution only from dark matter simulations, with significantly better results than conventional semi-analytical models (Wadekar et al., 2020). Hence, a HI distribution could be generated with this method in much larger volumes; even if is not as realistic as a proper hydrodynamic simulation, for a preliminary forecast such as this study it could be enough.

On the other hand, foregrounds should also be modelled in much larger areas, and while mock catalogues that simulate entire broad-band surveys exist, they are limited to the maximum magnitudes of said survey. Since the foregrounds that cannot be removed (and thus, need to be modelled) are actually the objects that cannot be resolved in the survey images (i.e., too faint for detection), any mock catalogue with the aim of simulating an existing survey will not be deep enough for foregrounds simulation. Running the semi-analytical codes used to generate these mock catalogues over N-body simulations at fainter magnitudes and higher redshifts is by no means a trivial task, so the fastest option might be taking an already existing catalogue and offsetting its objects to fainter magnitudes and higher redshifts. Further work should be done to determine how to do this while respecting observational constraints (e.g., luminosity functions), but given that it is a simulation of a noise component and not the actual signal, a highly realistic model would not be needed either.

In conclusion, we can say that this thesis presents a simulation of Lyα IM between the narrow-band images of PAUS and the Lyα forest data from eBOSS and DESI. The results of this forecast are mostly negative: even with the most optimistic estimate of uncorrelated instrumental noise in PAUS images, and assuming an extension of observing time for current PAUS, the probability of a real detection is just of 15.3% at best. However, one of the most important caveats of cross-correlating the Lyα emission from photometric images, the large smoothing in redshift direction due to the width of the filters, has been properly modelled and shown to maximise detection probability (and thus SNR) in different regimes depending on the type of 2PCF. This opens the door to similar studies with broad-band data, since the much larger redshift smoothing is no longer an issue, and the higher SNR due to wider filters, as well as the much larger angular coverages, would improve the SNR of the resulting cross-correlation with the Lyα forest.

Appendix A

Cross-correlation code

In this appendix, the actual core of the cross-correlation code is shown, i.e., the lines of code that carry out the bulk of the 2PCF computation. This can be interpreted as the real-code analogue to the pseudo-code displayed in 1. In fact, the code shown here would start at the same point, and end around 29, right before computing the 2PCF from the histograms.

Of course, this is not a fully functional correlation code on itself: the sections regarding data loading and management, as well as plotting of results have been omitted, as they are far more trivial and do not need any particular optimisation. Therefore, for code in this appendix to work, the following variables, directly derived from entry parameters of the code, are already expected to be defined:

- `n_bins_monopole`: Number of uniform distance bins for the monopole 2PCF.
- `n_bins_parallel`: Number of uniform distance bins for the parallel 2PCF.
- `n_bins_perpendicular`: Number of uniform distance bins for the perpendicular 2PCF.
- `parallel_cells`: Number of adjacent cells in parallel direction to compute cell pairs (size of PAUS cube surrounding forest cell).
- `perpendicular_cells`: Number of adjacent cells in perpendicular direction to compute cell pairs (size of PAUS cube surrounding forest cell).
- `jackknives`: Number of jackknife divisions in each angular direction (RA, dec). The number of jackknife subsamples will be $\texttt{jackknives}^2$.
- `array_size`: Size of the arrays to store cell pair data. This is a soft limit to how much memory the code will use.

Besides, the following arrays with the data to be cross-correlated should also be loaded:

- `grid`: δ values of all PAUS cells. Shape: [number of RA bins, number of dec bins, number of PAUS filters]. Floats.
- `v_lines`: δ values for all forest cells. Empty redshift bins for a forest (either because redshift is higher than the quasars, or the Lyβ forest is already present) should contain NaN. Shape: [number of forests, number of fine redshift bins]. Floats.
- `jk_v_lines`: Jackknife indices of each forest cell. Shape: [number of forests, number of fine redshift bins]. Integers.

- `v_lines_ind`: Indices of the PAUS cell in the `grid` array containing the respective forest cell; this array is used as a pointer for each forest cell. Shape: [3, number of forests, number of fine redshift bins]. Integers.
- `grid_coord`: Cartesian coordinates of all PAUS cells. Shape: [3, number of RA bins, number of DEC bins, number of PAUS filters]. Floats.
- `v_lines_coord`: Cartesian coordinates of all PAUS cells. Shape: [3, number of forests, number of fine redshift bins]. Floats.

Here, with *fine redshift bins* we refer to the actual redshift bins of the hydrodynamic simulation, as they are the effective maximum resolution of our simulation, and with *PAUS filters* we refer to the wide redshift bins resulting from merging the original bins to simulate the filters in PAUS images.

Regarding the technical aspects of the code, it has been written in Python 3 (run in version 3.6.12), and makes use of the library NumPy (version 1.18.5). The library is expected to have been loaded as follows:

```
import numpy as np
```

```
#%% Initializations
#Histograms vectors initialized
np_total, dd_total=np.zeros([n_bins_monopole,jackknives**2]),np.zeros([n_bins_monopole,jackknives**2])
np_cube=np.zeros((n_bins_parallel, n_bins_perpendicular,jackknives**2))
dd_cube=np.zeros((n_bins_parallel, n_bins_perpendicular,jackknives**2))
jk_histogram=np.zeros(jackknives**2)

# Initialization of cell pair vectors
ov_products=np.zeros(array_size).astype('float32')
dist_parallel=np.zeros(array_size).astype('float32')
dist_x, dist_y=np.zeros(array_size).astype('float32'), np.zeros(array_size).astype('float32')
jk_indices=np.zeros(array_size).astype('int16')

#%% Defining necessary functions
#Calling closest grid cell
def closest_grid(i, j, *args):
    if args==():
        return (grid_coord[:,v_lines_ind[0,i,j], v_lines_ind[1,i,j],v_lines_ind[2,i,j]])
    else:
        return (grid_coord[:,v_lines_ind[0,i,j]+args[0][0], v_lines_ind[1,i,j]+args[0][1],v_lines_ind[2,i,j]+args[0][2]])

#Histogram calculation
def hist_calculation(dist_parallel, jk_indices, dist_perpendicular, dist_total, ov_products, bins_parallel,
                     bins_perpendicular, bins_total, mask):
    for i in range(jackknives**2):
        chosen_ind=jk_indices==i
        np_total[:,i]=np_total[:,i]+np.bincount(dist_total[chosen_ind*~mask],
        ↪  minlength=n_bins_monopole)[0:n_bins_monopole]
        dd_total[:,i]=dd_total[:,i]+np.bincount(dist_total[chosen_ind*~mask], weights=ov_products[chosen_ind*~mask],
                                                minlength=n_bins_monopole)[0:n_bins_monopole]
        for j in range(n_bins_parallel):
            chosen_ind=np.logical_and(jk_indices==i, dist_parallel==j)
            np_cube[j,:,i]=np_cube[j,:,i]+np.bincount(dist_perpendicular[chosen_ind*~mask],
                                                      minlength=n_bins_perpendicular)[0:n_bins_perpendicular]
            dd_cube[j,:,i]=dd_cube[j,:,i]+np.bincount(dist_perpendicular[chosen_ind*~mask],
                    weights=ov_products[chosen_ind*~mask], minlength=n_bins_perpendicular)[0:n_bins_perpendicular]
    return(np_total,dd_total,np_cube, dd_cube, jk_histogram)
```

```
#%% Computation of distances between redshift planes and adjacent cells in redshift planes.

adjacent_parallel, adjacent_x, adjacent_y=np.zeros(grid.shape[2]-1), np.zeros(grid.shape[1:]), np.zeros(grid.shape[2])
for i in range(grid.shape[2]):
    adjacent_x[:,i]=np.linalg.norm(grid_coord[:,0,:,i]-grid_coord[:,1,:,i], axis=0)
    adjacent_y[i]=np.linalg.norm(grid_coord[:,0,0,i]-grid_coord[:,0,1,i])
    if i<grid.shape[2]-1:
        adjacent_parallel[i]=np.linalg.norm(grid_coord[:,0,0,i+1]-grid_coord[:,0,0,i])

# Calculation of scale factors for np.bincount
dist_max_par=(parallel_cells+1/2)*max(adjacent_parallel)
dist_max_perp=(perpendicular_cells+1/2)*np.linalg.norm((np.max(adjacent_x), np.max(adjacent_y)))
dist_max_mono=np.linalg.norm((dist_max_par, dist_max_perp))

scale_par=n_bins_parallel/dist_max_par
scale_perp=n_bins_perpendicular/dist_max_perp
scale_mono=n_bins_monopole/dist_max_mono

# Calculation of bin edges for 2D histogram
bins_parallel=np.arange(0, n_bins_parallel+1)
bins_perpendicular=np.arange(0, n_bins_perpendicular+1)
bins_total=np.arange(0, n_bins_monopole+1)

#%% Computing all cell pairs
# Initializing counter of number of cell pairs
n=0
```

```
for i in v_lines.shape[0]: #Iteration over vision lines

    # Obtaining maximum and minimum index in vision line (bin at quasar redshift)
    z_ind_min=0
    z_ind_max=v_lines.shape[1]
    for j in range(v_lines.shape[1]):
        if ~np.isnan(v_lines[i,j]):
            z_ind_min=j
            break
    for j in range(z_ind_min, v_lines.shape[1]):
        if np.isnan(v_lines[i,j]):
            z_ind_max=j
            break

    # Computation of unitary vectors perpendicular (line of sight) and paralel (x,y) to redshift plane
    # Only if the vision line has non-nan cells
    if z_ind_max>z_ind_min:
        if v_lines_ind[2,i,z_ind_min]<grid.shape[2]-1:
            axis_los=(closest_grid(i,z_ind_min,(0,0,1))-closest_grid(i,z_ind_min))/\
            np.linalg.norm(closest_grid(i,z_ind_min,(0,0,1))-closest_grid(i,z_ind_min))
        else:
            axis_los=(closest_grid(i,z_ind_min,(0,0,-1))-closest_grid(i,z_ind_min))/\
            np.linalg.norm(closest_grid(i,z_ind_min,(0,0,-1))-closest_grid(i,z_ind_min))
        if v_lines_ind[0,i,z_ind_min]<grid.shape[0]-1:
            axis_x=(closest_grid(i,z_ind_min,(1,0,0))-closest_grid(i,z_ind_min))/\
            np.linalg.norm(closest_grid(i,z_ind_min,(1,0,0))-closest_grid(i,z_ind_min))
        else:
            axis_x=(closest_grid(i,z_ind_min)-closest_grid(i,z_ind_min,(-1,0,0)))/\
            np.linalg.norm(closest_grid(i,z_ind_min)-closest_grid(i,z_ind_min,(-1,0,0)))
        if v_lines_ind[1,i,z_ind_min]<grid.shape[1]-1:
            axis_y=(closest_grid(i,z_ind_min,(0,1,0))-closest_grid(i,z_ind_min))/\
            np.linalg.norm(closest_grid(i,z_ind_min,(0,1,0))-closest_grid(i,z_ind_min))
        else:
            axis_y=(closest_grid(i,z_ind_min)-closest_grid(i,z_ind_min,(0,-1,0)))/\
            np.linalg.norm(closest_grid(i,z_ind_min)-closest_grid(i,z_ind_min,(0,-1,0)))
```

```
    for j in range(z_ind_min, z_ind_max): #Iteration over cells in vision line

        # Indices of the closest grid cell redefined to make things slightly faster. Also, it is checked if
        # these indices have not changed, in  order to avoid redundant calculations
        if j>z_ind_min:
            if ra_ind==v_lines_ind[0,i,j] and dec_ind==v_lines_ind[1,i,j] and z_ind==v_lines_ind[2,i,j]:
                same_grid_cell=True
            else:
                ra_ind=v_lines_ind[0,i,j]
                dec_ind=v_lines_ind[1,i,j]
                z_ind=v_lines_ind[2,i,j]
                same_grid_cell=False

            #Limits for computing correlations in adjacent redshift planes
            min_plane=max(z_ind-parallel_cells,0)
            max_plane=min(z_ind+parallel_cells+1,grid.shape[2])

            # Number of cell pairs for the vision line cell
            n_pairs_plane=displacements_x.shape[0]
            n_pairs_total=abs(max_plane-min_plane)*n_pairs_plane

        else:
            ra_ind=v_lines_ind[0,i,j]
            dec_ind=v_lines_ind[1,i,j]
            z_ind=v_lines_ind[2,i,j]
            same_grid_cell=False

            #Limits for computing correlations in adjacent redshift planes
            min_plane=max(z_ind-parallel_cells,0)
            max_plane=min(z_ind+parallel_cells+1,grid.shape[2])

            # Number of cell pairs for the vision line cell
            n_pairs_plane=displacements_x.shape[0]
            n_pairs_total=abs(max_plane-min_plane)*n_pairs_plane
```

```
        # Checking if cube of cell pairs does not fit in current storage arrays
        if n+n_pairs_total>=array_size: # If it does not fit, histograms are computed, storage arrays are emptied
            dist_perpendicular=np.linalg.norm((dist_x[:n], dist_y[:n]),axis=0)
            del dist_x
            del dist_y
            dist_total=np.linalg.norm((dist_parallel[:n], dist_perpendicular),axis=0)
            mask=np.isnan(ov_products)
            dist_parallel, dist_perpendicular=scale_par*dist_parallel, scale_perp*dist_perpendicular
            dist_total=scale_mono*dist_total
            dist_parallel, dist_perpendicular, dist_total=dist_parallel.astype('uint16'),
             ↪  dist_perpendicular.astype('uint16')
            dist_total=dist_total.astype('uint16')
            np_total,dd_total,np_cube, dd_cube, jk_histogram=hist_calculation(dist_parallel[:n], jk_indices[:n],
                dist_perpendicular, dist_total, ov_products[:n], bins_parallel, bins_perpendicular, bins_total, mask[:n])
            ov_products=np.zeros(array_size).astype('float32')
            dist_parallel=np.zeros(array_size).astype('float32')
            dist_x, dist_y=np.zeros(array_size).astype('float32'), np.zeros(array_size).astype('float32')
            n=0
            same_grid_cell=False
            ra_ind=v_lines_ind[0,i,j]
            dec_ind=v_lines_ind[1,i,j]
            z_ind=v_lines_ind[2,i,j]
```

```
        if same_grid_cell==False:
            #Computation of distance between actual vision line cell and closest grid cell (r_los, r_x, r_y)
            r=grid_coord[:,v_lines_ind[0,i,j], v_lines_ind[1,i,j],v_lines_ind[2,i,j]]-v_lines_coord[:,i,j]
            r_los=np.abs(np.dot(axis_los,r))
            r_x=np.dot(axis_x,r)
            r_y=np.dot(axis_y,r)

            # Distances and overdensity products for all surrounding grid cells computed
            dist_x_temp=np.repeat(adjacent_x[dec_ind, min_plane:max_plane], n_pairs_plane)\
                *np.tile(displacements_x, max_plane-min_plane)
            dist_y_temp=np.repeat(adjacent_y[min_plane:max_plane], n_pairs_plane)\
                *np.tile(displacements_y, max_plane-min_plane)
            dist_x[n:n+n_pairs_total]=dist_x_temp+r_x
            dist_y[n:n+n_pairs_total]=dist_y_temp+r_y
            back_planes_dist=np.cumsum(adjacent_parallel[min_plane:z_ind][::-1])[::-1]
            front_planes_dist=np.cumsum(adjacent_parallel[z_ind:max_plane-1])
            dist_r_los_temp=np.concatenate([-np.ones(((back_planes_dist.shape[0]+1)*n_pairs_plane)),
                                            np.ones((front_planes_dist.shape[0]*n_pairs_plane))])
            dist_parallel_temp=np.repeat(np.concatenate([back_planes_dist, [0], front_planes_dist]), n_pairs_plane)
            dist_parallel[n:n+n_pairs_total]=np.abs(dist_parallel_temp+r_los*dist_r_los_temp)
            adjacent_cells=np.reshape(grid[ra_ind+iteration_x[0]:ra_ind+iteration_x[-1]+1,
               dec_ind+iteration_y[0]:dec_ind+iteration_y[-1]+1, min_plane:max_plane], -1, order='F')
            ov_products[n:n+n_pairs_total]=adjacent_cells*v_lines[i,j]
            jk_indices[n:n+n_pairs_total]=jk_v_lines[i,j]*np.ones(n_pairs_total)
```

```
        else:
            # Storing distances between previous vision line cell and closest grid cell as other variables
            r=grid_coord[:,v_lines_ind[0,i,j], v_lines_ind[1,i,j],v_lines_ind[2,i,j]]-v_lines_coord[:,i,j]
            r_los=np.abs(np.dot(axis_los,r))
            r_x=np.dot(axis_x,r)
            r_y=np.dot(axis_y,r)

            # Distances and overdensity products computed from the previous vision line cell
            dist_x[n:n+n_pairs_total]=dist_x_temp+r_x
            dist_y[n:n+n_pairs_total]=dist_y_temp+r_y
            dist_parallel[n:n+n_pairs_total]=np.abs(dist_parallel_temp+r_los*dist_r_los_temp)
            ov_products[n:n+n_pairs_total]=adjacent_cells*v_lines[i,j]
            jk_indices[n:n+n_pairs_total]=jk_indices[n-n_pairs_total:n]

        n=n+n_pairs_total


#One last histogram calculation, if necessary.
dist_perpendicular=np.linalg.norm((dist_x[:n], dist_y[:n]),axis=0)
del dist_x
del dist_y
dist_total=np.linalg.norm((dist_parallel[:n], dist_perpendicular),axis=0)
mask=np.isnan(ov_products)
dist_parallel, dist_perpendicular=scale_par*dist_parallel, scale_perp*dist_perpendicular
dist_total=scale_mono*dist_total
dist_parallel, dist_perpendicular, dist_total=dist_parallel.astype('uint16'), dist_perpendicular.astype('uint16')
dist_total=dist_total.astype('uint16')
np_total,dd_total,np_cube, dd_cube, jk_histogram=hist_calculation(dist_parallel[:n], jk_indices[:n],
    dist_perpendicular, dist_total, ov_products[:n], bins_parallel, bins_perpendicular, bins_total, mask[:n])
```

Bibliography

Aarseth, S. J. and S. M. Fall (Feb. 1980). "Cosmological N-body simulations of galaxy merging". In: *Astrophys. J.* 236, p. 43. ISSN: 0004-637X. DOI: 10.1086/157716. URL: https://ui.adsabs.harvard.edu/abs/1980ApJ...236...43A/abstract.

Aarseth, S. J., E. L. Turner, and III Gott J. R. (Mar. 1979). "N-body simulations of galaxy clustering. I - Initial conditions and galaxy collapse times". In: *Astrophys. J.* 228, p. 664. ISSN: 0004-637X. DOI: 10.1086/156892. URL: https://ui.adsabs.harvard.edu/abs/1979ApJ...228..664A/abstract.

Abbott, B. P. et al. (Feb. 2016). "Observation of gravitational waves from a binary black hole merger". In: *Phys. Rev. Lett.* 116.6, p. 061102. ISSN: 10797114. DOI: 10.1103/PhysRevLett.116.061102. arXiv: 1602.03837. URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.116.061102.

Abbott, T. M. C. et al. (Jan. 2018). "The Dark Energy Survey Data Release 1". In: *Astrophys. J. Suppl. Ser.* 239.2, p. 18. DOI: 10.3847/1538-4365/aae9f0. arXiv: 1801.03181. URL: http://arxiv.org/abs/1801.03181http://dx.doi.org/10.3847/1538-4365/aae9f0.

Ahumada, Romina et al. (June 2020). "The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra". In: *Astrophys. J. Suppl. Ser.* 249.1, p. 3. ISSN: 1538-4365. DOI: 10.3847/1538-4365/ab929e. arXiv: 1912.02905. URL: https://arxiv.org/abs/1912.02905v2.

Alarcon, Alex et al. (July 2020). "The PAU Survey: An improved photo-z sample in the COSMOS field". In: arXiv: 2007.11132. URL: http://arxiv.org/abs/2007.11132.

Albrecht, Andreas and Paul J. Steinhardt (Apr. 1982). "Cosmology for grand unified theories with radiatively induced symmetry breaking". In: *Phys. Rev. Lett.* 48.17, pp. 1220–1223. ISSN: 00319007. DOI: 10.1103/PhysRevLett.48.1220. URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.48.1220.

Angulo, R. E. et al. (Nov. 2012). "Scaling relations for galaxy clusters in the millennium-XXL simulation". In: *Mon. Not. R. Astron. Soc.* 426.3, pp. 2046–2062. ISSN: 00358711. DOI: 10.1111/j.1365-2966.2012.21830.x. arXiv: 1203.3216. URL: https://academic.oup.com/mnras/article/426/3/2046/988710.

Bautista, Julian E. et al. (July 2017). "Measurement of baryon acoustic oscillation correlations at z = 2.3 with SDSS DR12 Ly-alpha Forests". In: 603, A12. DOI: 10.1051/0004-6361/201730533. arXiv: 1702.00176. URL: http://www.aanda.org/10.1051/0004-6361/201730533.

Becker, Robert H. et al. (Aug. 2001). "Evidence for Reionization at z$\sim$6: Detection of a Gunn-Peterson Trough in a z=6.28 Quasar". In: *Astron. J.* 122.6, pp. 2850–2857. DOI: 10.1086/324231. arXiv: 0108097 [astro-ph]. URL: http://arxiv.org/abs/astro-ph/0108097http://dx.doi.org/10.1086/324231.

Benson, Andrew J. (Feb. 2012). "Galacticus: A semi-analytic model of galaxy formation". In: *New Astron.* 17.2, pp. 175–197. ISSN: 13841076. DOI: 10.1016/j.newast.2011.07.004. arXiv: 1008.1786.

Bernal, Jose Luis, Licia Verde, and Adam G. Riess (July 2016). "The trouble with H_0". In: *J. Cosmol. Astropart. Phys.* DOI: 10.1088/1475-7516/2016/10/019. arXiv: 1607.05617. URL: http://arxiv.org/abs/1607.05617http://dx.doi.org/10.1088/1475-7516/2016/10/019.

Bernardeau, F. et al. (Dec. 2001). "Large-Scale Structure of the Universe and Cosmological Perturbation Theory". In: *Phys. Rep.* 367.1-3, pp. 1–248. DOI: 10.1016/S0370-1573(02)00135-7. arXiv: 0112551 [astro-ph]. URL: http://arxiv.org/abs/astro-ph/0112551http://dx.doi.org/10.1016/S0370-1573(02)00135-7.

Bertin, Emmanuel et al. (2002). "The TERAPIX Pipeline". In: *Astron. Data Anal. Softw. Syst. XI*. Ed. by David Bohlender, Daniel Durand, and Thomas Handley. San Francisco: Astronomical Society of the Pacific, p. 228. ISBN: 1-58381-124-9. URL: https://ui.adsabs.harvard.edu/abs/2002ASPC..281..228B/abstract.

Blanton, Michael R. and Sam Roweis (Feb. 2006). "K-corrections and filter transformations in the ultraviolet, optical, and near infrared". In: *Astron. J.* 133.2, pp. 734–755. ISSN: 0004-6256. DOI: 10.1086/510127. arXiv: 0606170 [astro-ph]. URL: http://stacks.iop.org/1538-3881/133/i=2/a=734http://arxiv.org/abs/astro-ph/0606170{\%}0Ahttp://dx.doi.org/10.1086/510127.

Blomqvist, Michael et al. (Sept. 2019). " Baryon acoustic oscillations from the cross-correlation of Ly α absorption and quasars in eBOSS DR14 ". In: *Astron. Astrophys.* 629, A86. ISSN: 0004-6361. DOI: 10.1051/0004-6361/201935641. arXiv: 1904.03430.

Breysse, Patrick C., Ely D. Kovetz, and Marc Kamionkowski (Oct. 2014). "Carbon monoxide intensity mapping at moderate redshifts". In: *Mon. Not. R. Astron. Soc.* 443.4, pp. 3506–3512. ISSN: 1365-2966. DOI: 10.1093/mnras/stu1312. URL: http://academic.oup.com/mnras/article/443/4/3506/1008614/Carbon-monoxide-intensity-mapping-at-moderate.

Carilli, C. L. (Apr. 2011). "Intensity Mapping of molecular gas during cosmic reionization". In: *Astrophys. J.* 730.2, p. L30. ISSN: 2041-8205. DOI: 10.1088/2041-8205/730/2/L30. URL: http://stacks.iop.org/2041-8205/730/i=2/a=L30?key=crossref.f88195fbf912542448f1834c7ee6768b.

Carswell, R. F. et al. (Jan. 1975). "The spectrum of OH 471 /0642+44/". In: *Astrophys. J.* 195, p. 269. ISSN: 0004-637X. DOI: 10.1086/153326. URL: https://ui.adsabs.harvard.edu/abs/1975ApJ...195..269C/abstract.

Carswell, R. F. et al. (Aug. 1987). "High-redshift QSO absorbing clouds and the background ionizing source". In: *Astrophys. J.* 319, p. 709. ISSN: 0004-637X. DOI: 10.1086/165491. URL: https://ui.adsabs.harvard.edu/abs/1987ApJ...319..709C/abstract.

Cassata, P. et al. (Jan. 2011). "The VIMOS VLT Deep Survey: star formation rate density of Ly α emitters from a sample of 217 galaxies with spectroscopic redshifts $2 \leq z \leq 6.6$". In: *Astron. Astrophys.* 525, A143. ISSN: 0004-6361. DOI: 10.1051/0004-6361/201014410. URL: http://www.aanda.org/10.1051/0004-6361/201014410.

Castander, Francisco J. et al. (Sept. 2012). "The PAU camera and the PAU survey at the William Herschel Telescope". In: ed. by Ian S. McLean, Suzanne K. Ramsay, and Hideki Takami. Vol. 8446. International Society for Optics and Photonics, p. 84466D. DOI: 10.1117/12.926234. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.926234.

Chabanier, Solène et al. (2019). "The one-dimensional power spectrum from the SDSS DR14 Ly-alpha forests". In: *J. Cosmol. Astropart. Phys.* 7. ISSN: 14757516. DOI: 10.1088/1475-7516/2019/07/017. arXiv: 1812.03554. URL: http://arxiv.org/abs/1812.03554http://dx.doi.org/10.1088/1475-7516/2019/07/017.

Chang, Tzu-ching et al. (Mar. 2008). "Baryon Acoustic Oscillation Intensity Mapping as a Test of Dark Energy". In: *Phys. Rev. Lett.* 100.9, pp. 1–5. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.100.091303. arXiv: arXiv:0709.3672v2. URL: https://link.aps.org/doi/10.1103/PhysRevLett.100.091303.

Chang, Tzu-Ching et al. (July 2010). "An intensity map of hydrogen 21-cm emission at redshift $z \approx 0.8$". In: *Nature* 466.7305, pp. 463–465. ISSN: 0028-0836. DOI: 10.1038/nature09187. URL: http://www.nature.com/articles/nature09187.

Chiang, Yi-Kuan, Brice Ménard, and David Schiminovich (Oct. 2019). "Broadband Intensity Tomography: Spectral Tagging of the Cosmic UV Background". In: *Astrophys. J.* 877.2, p. 150. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ab1b35. arXiv: 1810.00885. URL: http://arxiv.org/abs/1810.00885http://dx.doi.org/10.3847/1538-4357/ab1b35.

Cisewski, Jessi et al. (May 2014). "Non-parametric 3D map of the intergalactic medium using the Lyman-alpha forest". In: *Mon. Not. R. Astron. Soc.* 440.3, pp. 2599–2609. ISSN: 0035-8711. DOI: 10.1093/mnras/stu475. URL: https://academic.oup.com/mnras/article/440/3/2599/1749970.

Cole, Shaun et al. (Sept. 2005). "The 2dF Galaxy Redshift Survey: Power-spectrum analysis of the final data set and cosmological implications". In: 362.2, pp. 505–534. ISSN: 00358711. DOI: 10.1111/j.1365-2966.2005.09318.x. arXiv: 0501174 [astro-ph]. URL: https://academic.oup.com/mnras/article/362/2/505/1017047.

Croft, Rupert A. C. et al. (2018). "Intensity mapping with SDSS/BOSS Lyman-alpha emission, quasars and their Lyman-alpha forest". In: *Mon. Not. R. Astron. Soc.* 000.June. DOI: 10.1093/mnras/sty2302. arXiv: 1806.06050. URL: http://arxiv.org/abs/1806.06050{\%}0Ahttp://dx.doi.org/10.1093/mnras/sty2302.

Croft, Rupert A.C. et al. (2016). "Large-scale clustering of Lyman α emission intensity from SDSS/BOSS". In: *Mon. Not. R. Astron. Soc.* 457.4, pp. 3541–3572. ISSN: 13652966. DOI: 10.1093/mnras/stw204. arXiv: 1504.04088.

Davis, M., J. Huchra, and D. Latham (1983). "The center for astrophysics redshift survey". In: *IAUS* 104, pp. 167–172. ISSN: 1743-9221. URL: https://ui.adsabs.harvard.edu/abs/1983IAUS..104..167D/abstract.

Davis, Tamara M. and Charles H. Lineweaver (Nov. 2001). "Superluminal recession velocities". In: *AIPC*. Vol. 555. AIP Publishing, pp. 348–351. DOI: 10.1063/1.1363540. arXiv: 0011070 [astro-ph]. URL: http://arxiv.org/abs/astro-ph/0011070http://dx.doi.org/10.1063/1.1363540https://ui.adsabs.harvard.edu/abs/2001AIPC..555..348D/abstract.

Dawson, Kyle S. et al. (Jan. 2013). "The Baryon Oscillation Spectroscopic Survey of SDSS-III". In: *Astron. J.* 145.1, p. 10. ISSN: 0004-6256. DOI: 10.1088/0004-6256/145/1/10. URL: http://stacks.iop.org/1538-3881/145/i=1/a=10?key=crossref.af65e7d4595bf6350e8598f97f044dc6.

Dawson, Kyle S. et al. (Feb. 2016). "The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and early data". In: *Astron. J.* 151.2, p. 44. ISSN: 1538-3881. DOI: 10.3847/0004-6256/151/2/44. URL: http://stacks.iop.org/1538-3881/151/i=2/a=44?key=crossref.239d6a216cbcb1fd1f791f9dc1674011.

de Sainte Agathe, Victoria et al. (Apr. 2019). "Baryon acoustic oscillations at z = 2.34 from the correlations of Lyα absorption in eBOSS DR14". In: *Astronomy & Astrophysics* 629, A85. ISSN: 0004-6361. DOI: 10.1051/0004-6361/201935638. arXiv: 1904.03400. URL: http://arxiv.org/abs/1904.03400.

DESI Collaboration et al. (Oct. 2016a). *The DESI Experiment Part II: Instrument Design*. Tech. rep. arXiv: 1611.00037. URL: http://arxiv.org/abs/1611.00037.

DESI Collaboration, DESI et al. (Oct. 2016b). *The DESI Experiment Part I: Science,Targeting, and Survey Design*. Tech. rep. arXiv: 1611.00036. URL: http://arxiv.org/abs/1611.00036.

Desjacques, Vincent, Donghui Jeong, and Fabian Schmidt (Feb. 2018). "Large-scale galaxy bias". In: *Phys. Rep.* 733, pp. 1–193. ISSN: 03701573. DOI: 10.1016/j.physrep.2017.12.002. URL: https://ui.adsabs.harvard.edu/abs/2018PhR...733....1D/abstract.

Di Matteo, T. et al. (Feb. 2012). "Cold flows and the first quasars". In: *Astrophys. J.* 745.2, p. L29. ISSN: 2041-8205. DOI: 10.1088/2041-8205/745/2/L29. URL: http://stacks.iop.org/2041-8205/745/i=2/a=L29?key=crossref.ee3351a31752a1906e05effa3f33aff9.

Dodelson, Scott and Fabian Schmidt (2020). *Modern Cosmology (Second Edition)*. Ed. by Scott Dodelson and Fabian Schmidt. 2nd Edition. Academic Press. ISBN: 9780128159484. DOI: https://doi.org/10.1016/B978-0-12-815948-4.00006-1..

Einstein, Albert (1915). "Die Feldgleichungen der Gravitation". In: *Sitzungsberichte der Königlich Preußischen Akad. der Wissenschaften*, pp. 844–847.

— (1917). "Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie". In: *Sitzungsberichte der Königlich Preußischen Akad. der Wissenschaften*, pp. 142–152.

Eisenstein, Daniel J. et al. (Nov. 2005). "Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies". In: *Astrophys. J.* 633.2, pp. 560–574. ISSN: 0004-637X. DOI: 10.1086/466512. arXiv: 0501171 [astro-ph]. URL: https://iopscience.iop.org/article/10.1086/466512https://iopscience.iop.org/article/10.1086/466512/meta.

Eisenstein, Daniel J. et al. (Sept. 2011). "SDSS-III: Massive spectroscopic surveys of the distant Universe, the Milky Way, and extra-solar planetary systems". In: *Astron. J.* 142.3, p. 72. ISSN: 0004-6256. DOI: 10.1088/0004-6256/142/3/72. URL: http://stacks.iop.org/1538-3881/142/i=3/a=72?key=crossref.c4bf9c63075c50b5eb07ebbbf2452a05.

Eriksen, M. et al. (Apr. 2019). "The PAU survey: Early demonstration of photometric redshift performance in the COSMOS field". In: *Mon. Not. R. Astron. Soc.* 484.3, pp. 4200–4215. ISSN: 13652966. DOI: 10.1093/mnras/stz204. arXiv: 1809.04375.

Escoffier, S. et al. (2016). *Jackknife resampling technique on mocks: an alternative method for covariance matrix estimation*. arXiv: 1606.00233. URL: http://arxiv.org/abs/1606.00233.

Ezquiaga, Jose María and Miguel Zumalacárregui (Dec. 2018). "Dark Energy in Light of Multi-Messenger Gravitational-Wave Astronomy". In: *Front. Astron. Sp. Sci.* 5, p. 44. ISSN: 2296987X. DOI: 10.3389/fspas.2018.00044. arXiv: 1807.09241. URL: www.frontiersin.org.

Font-Ribera, Andreu et al. (Sept. 2012). "The large-scale cross-correlation of Damped Lyman Alpha Systems with the Lyman Alpha Forest: First Measurements from BOSS". In: *J. Cosmol. Astropart. Phys.* 11. ISSN: 14757516. DOI: 10.1088/1475-7516/2012/11/059. arXiv: 1209.4596. URL: http://arxiv.org/abs/1209.4596http://dx.doi.org/10.1088/1475-7516/2012/11/059.

Fosalba, P. et al. (Dec. 2013). "The MICE Grand Challenge Lightcone Simulation I: Dark matter clustering". In: *Mon. Not. R. Astron. Soc.* 448.4, pp. 2987–3000. DOI: 10.1093/mnras/stv138. arXiv: 1312.1707. URL: http://arxiv.org/abs/1312.1707http://dx.doi.org/10.1093/mnras/stv138.

Friedmann, A. (Dec. 1922). "Über die Krümmung des Raumes". In: *Zeitschrift für Phys.* 10.1, pp. 377–386. ISSN: 14346001. DOI: 10.1007/BF01332580. URL: https://link.springer.com/article/10.1007/BF01332580.

— (Dec. 1924). "Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes". In: *Zeitschrift für Phys.* 21.1, pp. 326–332. ISSN: 14346001. DOI: 10.1007/BF01328280. URL: https://link.springer.com/article/10.1007/BF01328280.

Garzilli, Antonella, Alexey Boyarsky, and Oleg Ruchayskiy (Oct. 2017). "Cutoff in the Lyman-α forest power spectrum: Warm IGM or warm dark matter?" In: *Phys. Lett. Sect. B Nucl. Elem. Part. High-Energy Phys.* 773, pp. 258–264. ISSN: 03702693. DOI: 10.1016/j.physletb.2017.08.022. arXiv: 1510.07006. URL: https://ui.adsabs.harvard.edu/abs/2017PhLB..773..258G/abstract.

Gawiser, Eric et al. (Dec. 2007). "Lyα-Emitting Galaxies at z = 3.1: L* Progenitors Experiencing Rapid Star Formation". In: *Astrophys. J.* 671.1, pp. 278–284. DOI: 10.1086/522955. arXiv: 0710.2697 [astro-ph].

Gaztañaga, Enrique, Anna Cabré, and Lam Hui (Nov. 2009). "Clustering of luminous red galaxies - IV. Baryon acoustic peak in the line-of-sight direction and a direct measurement of H(z)". In: *Mon. Not. R. Astron. Soc.* 399.3, pp. 1663–1680. DOI: 10.1111/j.1365-2966.2009.15405.x. arXiv: 0807.3551 [astro-ph].

Gaztañaga, Enrique et al. (June 2012). "Cross-correlation of spectroscopic and photometric galaxy surveys: cosmology from lensing and redshift distortions". In: *Mon. Not. R. Astron. Soc.* 422.4, pp. 2904–2930. ISSN: 00358711. DOI: 10.1111/j.1365-2966.2012.20613.x. URL: https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2012.20613.x.

Gil de Paz, Armando et al. (Dec. 2007). " The GALEX Ultraviolet Atlas of Nearby Galaxies ". In: *Astrophys. J. Suppl. Ser.* 173.2, pp. 185–255. ISSN: 0067-0049. DOI: 10.1086/516636. arXiv: 0606440 [astro-ph].

Gingold, R. A. and J. J. Monaghan (Dec. 1977). "Smoothed particle hydrodynamics: theory and application to non-spherical stars". In: *Mon. Not. R. Astron. Soc.* 181.3, pp. 375–389. ISSN: 0035-8711. DOI: 10.1093/mnras/181.3.375. URL: https://academic.oup.com/mnras/article/181/3/375/988212.

Gonçalves, R S et al. (Dec. 2018). "Measuring the scale of cosmic homogeneity with SDSS-IV DR14 quasars". In: *Mon. Not. R. Astron. Soc.* 481.4, pp. 5270–5274. ISSN: 0035-8711. DOI: 10.1093/mnras/sty2670. URL: https://academic.oup.com/mnras/article/481/4/5270/5114592.

Gong, Yan et al. (Jan. 2012). "Intensity Mapping of the [C II] fine structure line during the epoch of reionization". In: *Astrophys. J.* 745.1, p. 49. ISSN: 0004-637X. DOI: 10.1088/0004-637X/745/1/49. URL: http://stacks.iop.org/0004-637X/745/i=1/a=49?key=crossref.43e164ebef42db5c144862887f42f97c.

Gonzalez-perez, V. et al. (Mar. 2014). "How sensitive are predicted galaxy luminosities to the choice of stellar population synthesis model?" In: *Mon. Not. R. Astron. Soc.* 439.1, pp. 264–283. ISSN: 00358711. DOI: 10.1093/mnras/stt2410.

Gronwall, Caryl et al. (Sept. 2007). " Lyα Emission-Line Galaxies at z = 3.1 in the Extended Chandra Deep Field–South ". In: *Astrophys. J.* 667.1, pp. 79–91. ISSN: 0004-637X. DOI: 10.1086/520324. arXiv: 0705.3917.

Gunn, James E. and Bruce A. Peterson (Nov. 1965). "On the Density of Neutral Hydrogen in Intergalactic Space." In: *Astrophys. J.* 142, p. 1633. ISSN: 0004-637X. DOI: 10.1086/148444. URL: https://ui.adsabs.harvard.edu/abs/1965ApJ...142.1633G/abstract.

Guo, Qi et al. (2013). "Galaxy formation in WMAP1 and WMAP7 cosmologies". In: *Mon. Not. R. Astron. Soc.* 428.2, pp. 1351–1365. ISSN: 1365-2966. DOI: 10.1093/mnras/sts115. arXiv: 1206.0052. URL: http://arxiv.org/abs/1206.0052http://dx.doi.org/10.1093/mnras/sts115http://academic.oup.com/mnras/

article/428/2/1351/1004346/Galaxy-formation-in-WMAP1-and-WMAP7-cosmologies.

Guth, Alan H. and So Young Pi (Oct. 1982). "Fluctuations in the new inflationary universe". In: *Phys. Rev. Lett.* 49.15, pp. 1110–1113. ISSN: 00319007. DOI: 10.1103/PhysRevLett.49.1110. URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.49.1110.

Hatton, Steve et al. (July 2003). "GALLICS - I. A hybrid N-body/semi-analytic model of hierarchical galaxy formation". In: *Mon. Not. R. Astron. Soc.* 343.1, pp. 75–106. ISSN: 00358711. DOI: 10.1046/j.1365-8711.2003.05589.x. arXiv: 0309186 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/2003MNRAS.343...75H/abstract.

Hellwing, Wojciech A. et al. (Feb. 2016). "The copernicus complexio: A high-resolution view of the small-scale universe". In: *Mon. Not. R. Astron. Soc.* 457.4, pp. 3492–3509. ISSN: 13652966. DOI: 10.1093/mnras/stw214. arXiv: 1505.06436. URL: https://academic.oup.com/mnras/article/457/4/3492/2588996.

Hernquist, Lars et al. (Feb. 1996). "The Lyman-Alpha Forest in the Cold Dark Matter Model". In: *Astrophys. J. Letters* 457, p. L51. DOI: 10.1086/309899. arXiv: astro-ph/9509105 [astro-ph].

Heymans, Catherine et al. (2012). "CFHTLenS: The Canada-France-Hawaii Telescope Lensing Survey". In: *Mon. Not. R. Astron. Soc.* 427.1, pp. 146–166. ISSN: 00358711. DOI: 10.1111/j.1365-2966.2012.21952.x. arXiv: 1210.0032. URL: http://arxiv.org/abs/1210.0032http://dx.doi.org/10.1111/j.1365-2966.2012.21952.xhttps://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2012.21952.x.

Hoyle, B. et al. (July 2018). "Dark energy survey year 1 results: Redshift distributions of the weak-lensing source galaxies". In: *Mon. Not. R. Astron. Soc.* 478.1, pp. 592–610. ISSN: 13652966. DOI: 10.1093/mnras/sty957. arXiv: 1708.01532. URL: https://academic.oup.com/mnras/article/478/1/592/4975790.

Hoyle, F. (Oct. 1948). "A New Model for the Expanding Universe". In: *Mon. Not. R. Astron. Soc.* 108.5, pp. 372–382. ISSN: 0035-8711. DOI: 10.1093/mnras/108.5.372. URL: https://academic.oup.com/mnras/article/108/5/372/2601825.

Hubble, E. (Mar. 1929). "A relation between distance and radial velocity among extra-galactic nebulae". In: *Proc. Natl. Acad. Sci.* 15.3, pp. 168–173. ISSN: 0027-8424. DOI: 10.1073/pnas.15.3.168. URL: https://www.pnas.org/content/15/3/168https://www.pnas.org/content/15/3/168.abstract.

Hui, Lam, Enrique Gaztañaga, and Marilena Loverde (Nov. 2007). "Anisotropic magnification distortion of the 3D galaxy correlation. I. Real space". In: *Phys. Rev. D* 76.10, 103502, p. 103502. DOI: 10.1103/PhysRevD.76.103502. arXiv: 0706.1071 [astro-ph].

Jarvis, M., G. Bernstein, and B. Jain (July 2004). "The skewness of the aperture mass statistic". In: *Mon. Not. R. Astron. Soc.* 352.1, pp. 338–352. ISSN: 00358711. DOI: 10.1111/j.1365-2966.2004.07926.x. arXiv: 0307393 [astro-ph]. URL: https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2004.07926.x.

Kaiser, Nick (July 1987). "Clustering in real space and in redshift space". In: *Mon. Not. R. Astron. Soc.* 227.1, pp. 1–21. ISSN: 0035-8711. DOI: 10.1093/mnras/227.1.1. URL: https://academic.oup.com/mnras/article/227/1/1/1065830.

Kerscher, Martin, István Szapudi, and Alexander S. Szalay (May 2000). "A Comparison of Estimators for the Two-Point Correlation Function". In: *Astrophys. J.* 535.1, pp. L13–L16. ISSN: 0004637X. DOI: 10.1086/312702. arXiv: 9912088 [astro-ph].

Kovetz, E D et al. (2017). *Line-Intensity Mapping : 2017 Status Report*. Tech. rep. June. arXiv: arXiv:1709.09066v1. URL: https://arxiv.org/abs/1709.09066.

Laigle, C. et al. (June 2016). "The COSMOS2015 catalog: exploring the 1 < z < 6 universe with half a million galaxies". In: *Astrophys. J. Suppl. Ser.* 224.2, p. 24. ISSN: 1538-4365. DOI: 10.3847/0067-0049/224/2/24. URL: http://stacks.iop.org/0067-0049/224/i=2/a=24?key=crossref.13400bd5b9499f45ce2b1d44c4a48475.

Lemaître, Georges. In: *Publ. du Lab. d'Astronomie Geod. l'Universite Louvain* 12 ().

— (1927). "Un Univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques". In: *Ann. la Société Sci. Bruxelles, A47, p. 49-59* 47, p. 49. URL: https://ui.adsabs.harvard.edu/abs/1927ASSB...47...49L/abstract.

— (1931). "Expansion of the universe, A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae". In: *Mon. Not. R. Astron. Soc.* 91, pp. 483–490.

Lewis, Antony, Anthony Challinor, and Anthony Lasenby (Aug. 2000). "Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models". In: *Astrophys. J.* 538.2, pp. 473–476. ISSN: 0004-637X. DOI: 10.1086/309179.

Li, Tony Y. et al. (Jan. 2016). "Connecting CO Intensity Mapping to molecular gas and star formation in the epoch of galaxy assembly". In: *Astrophys. J.* 817.2, p. 169. ISSN: 1538-4357. DOI: 10.3847/0004-637X/817/2/169. arXiv: arXiv:1503.08833v2. URL: http://stacks.iop.org/0004-637X/817/i=2/a=169?key=crossref.1fc42cae7cb96f7a5fe6275ac9bc3cc9.

Libeskind, Noam I et al. (Sept. 2009). "Constrained simulations of the Local Group: on the radial distribution of substructures". In: *Mon. Not. R. Astron. Soc.* 401.3, pp. 1889–1897. DOI: 10.1111/j.1365-2966.2009.15766.x. arXiv: 0909.4423. URL: http://arxiv.org/abs/0909.4423http://dx.doi.org/10.1111/j.1365-2966.2009.15766.x.

Linde, A. D. (Feb. 1982). "A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems". In: *Phys. Lett. B* 108.6, pp. 389–393. ISSN: 03702693. DOI: 10.1016/0370-2693(82)91219-9.

Liu, Adrian et al. (Mar. 2009). "An Improved Method for 21cm Foreground Removal". In: *Mon. Not. R. Astron. Soc.* 398.1, pp. 401–406. DOI: 10.1111/j.1365-2966.2009.15156.x. arXiv: 0903.4890. URL: http://arxiv.org/abs/0903.4890http://dx.doi.org/10.1111/j.1365-2966.2009.15156.x.

Loeb, Abraham and Matias Zaldarriaga (May 2004). "Measuring the Small-Scale Power Spectrum of Cosmic Density Fluctuations Through 21 cm Tomography Prior to the Epoch of Structure Formation". In: *Phys. Rev. Lett.* 92.21, p. 211301. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.92.211301. arXiv: 0312134v2 [arXiv:astro-ph]. URL: https://link.aps.org/doi/10.1103/PhysRevLett.92.211301.

Lynds, Roger (Mar. 1971). "The Absorption-Line Spectrum of 4c 05.34". In: *Astrophys. J.* 164, p. L73. ISSN: 0004-637X. DOI: 10.1086/180695. URL: https://ui.adsabs.harvard.edu/abs/1971ApJ...164L..73L/abstract.

Madau, Piero et al. (Feb. 1997). "21-cm Tomography of the Intergalactic Medium at high redshift". In: *Astrophys. J.* 475.2, pp. 429–444. ISSN: 0004-637X. DOI: 10.1086/303549. arXiv: 9608010v1 [arXiv:astro-ph]. URL: http://stacks.iop.org/0004-637X/475/i=2/a=429.

Martí, P. et al. (July 2014). "Precise photometric redshifts with a narrow-band filter set: the PAU survey at the William Herschel Telescope". In: *Mon. Not. R. Astron. Soc.* 442.1, pp. 92–109. ISSN: 1365-2966. DOI: 10.1093/mnras/stu801. URL:

http://academic.oup.com/mnras/article/442/1/92/1229670/Precise-photometric-redshifts-with-a-narrowband.

Matsuda, Yuichi et al. (Aug. 2004). "A Subaru Search for Lyα Blobs in and around the Protocluster Region At Redshift z = 3.1". In: *Astron. J.* 128.2, pp. 569–584. DOI: 10.1086/422020. URL: http://stacks.iop.org/1538-3881/128/i=2/a=569.

Merson, Alexander I. et al. (2013). "Lightcone mock catalogues from semi-analytic models of galaxy formation - I. construction and application to the bzk colour selection". In: *Mon. Not. R. Astron. Soc.* 429.1, pp. 556–578. ISSN: 13652966. DOI: 10.1093/mnras/sts355. arXiv: 1206.4049.

Miller, A. D. et al. (Oct. 1999). "A Measurement of the Angular Power Spectrum of the Cosmic Microwave Background from l = 100 to 400". In: *Astrophys. J.* 524.1, pp. L1–L4. ISSN: 0004637X. DOI: 10.1086/312293. URL: http://www.hep.upenn.edu/CBR/.

O'Connell, Ross et al. (Aug. 2019). "Large Covariance Matrices: Accurate Models Without Mocks". In: *Mon. Not. R. Astron. Soc.* 487.2, pp. 2701–2717. ISSN: 0035-8711. DOI: 10.1093/mnras/stz1359. arXiv: 1808.05978. URL: http://arxiv.org/abs/1808.05978https://academic.oup.com/mnras/article/487/2/2701/5493113.

Ozbek, Melih, Rupert A. C. Croft, and Nishikanta Khandai (Feb. 2016). "Large-scale 3D mapping of the intergalactic medium using the Lyman Alpha Forest". In: *Mon. Not. R. Astron. Soc.* 456.4, pp. 3610–3623. DOI: 10.1093/mnras/stv2894. arXiv: 1602.01862. URL: http://arxiv.org/abs/1602.01862http://dx.doi.org/10.1093/mnras/stv2894.

Pandey, Biswajit and Suman Sarkar (Dec. 2015). "Testing homogeneity in the Sloan Digital Sky Survey Data Release Twelve with Shannon entropy". In: *Mon. Not. R. Astron. Soc.* 454.3, pp. 2647–2656. ISSN: 13652966. DOI: 10.1093/mnras/stv2166. arXiv: 1507.03124. URL: https://academic.oup.com/mnras/article/454/3/2647/1205693.

Peacock, J. A. et al. (2014). *HALOFIT: Nonlinear distribution of cosmological mass and galaxies*.

Peebles, P. J. E. and J. T. Yu (Dec. 1970). "Primeval Adiabatic Perturbation in an Expanding Universe". In: *Astrophys. J.* 162, p. 815. ISSN: 0004-637X. DOI: 10.1086/150713. URL: https://ui.adsabs.harvard.edu/abs/1970ApJ...162..815P/abstract.

Peebles, P.J.E. (1980). *The large-scale structure of the universe*. Princeton University Press. ISBN: 978-0-691-08240-0.

Peterson, Jeffrey B. et al. (2009). "21-cm Intensity Mapping". In: *astro2010 Astron. Astrophys. Decad. Surv.* 2010, p. 234. URL: https://ui.adsabs.harvard.edu/{\#}abs/2009astro2010S.234P/abstract.

Planck Collaboration (Apr. 2020). "Planck 2018 results. VI. Cosmological parameters". In: *Astron. Astrophys.* 63. ISSN: 0004-6361. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209. URL: https://pla.esac.esa.int.

Planck Collaboration et al. (Oct. 2016). "Planck 2015 results: XIII. Cosmological parameters". In: *Astron. Astrophys.* 594. ISSN: 14320746. DOI: 10.1051/0004-6361/201525830. arXiv: 1502.01589. URL: http://www.esa.int/Planck.

Potter, Douglas, Joachim Stadel, and Romain Teyssier (Dec. 2017). "PKDGRAV3: beyond trillion particle cosmological simulations for the next era of galaxy surveys". In: *Comput. Astrophys. Cosmol.* 4.1, p. 2. ISSN: 2197-7909. DOI: 10.1186/s40668-017-0021-1. arXiv: 1609.08621. URL: https://ui.adsabs.harvard.edu/abs/2017ComAC...4....2P/abstract.

Prochaska, Jason X. (Feb. 1999). "The Physical Nature of the Lyman-Limit Systems". In: *Astrophys. J.* 511.2, pp. L71–L74. ISSN: 0004637X. DOI: 10.1086/311849. URL: https://iopscience.iop.org/article/10.1086/311849https://iopscience.iop.org/article/10.1086/311849/meta.

Pullen, Anthony R., Olivier Doré, and Jamie Bock (Apr. 2014). "Intensity Mapping across cosmic times with the Lyα line". In: *Astrophys. J.* 786.2, p. 111. ISSN: 0004-637X. DOI: 10.1088/0004-637X/786/2/111. URL: http://stacks.iop.org/0004-637X/786/i=2/a=111?key=crossref.32d2f12627360598f3105e20c276339a.

Pullen, Anthony R et al. (Aug. 2018). "Search for C II emission on cosmological scales at redshift Z $\sim$ 2.6". In: *Mon. Not. R. Astron. Soc.* 478.2, pp. 1911–1924. ISSN: 0035-8711. DOI: 10.1093/mnras/sty1243. arXiv: arXiv:1707.06172v2. URL: https://academic.oup.com/mnras/article/478/2/1911/4995923.

Rauch, Michael (1998). "The Lyman Alpha Forest in the Spectra of QSOs". In: *Annu. Rev. Astron. Astrophys.* 36.1, pp. 267–316. ISSN: 0066-4146. DOI: 10.1146/annurev.astro.36.1.267. arXiv: 9806286 [astro-ph]. URL: http://www.annualreviews.org/doi/10.1146/annurev.astro.36.1.267http://arxiv.org/abs/astro-ph/9806286http://dx.doi.org/10.1146/annurev.astro.36.1.267.

Robertson, H. P. (Nov. 1935). "Kinematics and World-Structure". In: *Astrophys. J.* 82, p. 284. ISSN: 0004-637X. DOI: 10.1086/143681. URL: https://ui.adsabs.harvard.edu/abs/1935ApJ....82..284R/abstract.

Schaye, Joop et al. (Jan. 2015). "The EAGLE project: Simulating the evolution and assembly of galaxies and their environments". In: *Mon. Not. R. Astron. Soc.* 446.1, pp. 521–554. ISSN: 13652966. DOI: 10.1093/mnras/stu2058. arXiv: 1407.7040. URL: https://ui.adsabs.harvard.edu/abs/2015MNRAS.446..521S/abstract.

Schwarz, Dominik J et al. (Aug. 2016). "CMB anomalies after Planck". In: *Class. Quantum Gravity* 33.18, p. 184001. ISSN: 0264-9381. DOI: 10.1088/0264-9381/33/18/184001. URL: https://iopscience.iop.org/article/10.1088/0264-9381/33/18/184001https://iopscience.iop.org/article/10.1088/0264-9381/33/18/184001/meta.

Seo, Hee-Jong and Daniel J. Eisenstein (Dec. 2003). "Probing Dark Energy with Baryonic Acoustic Oscillations from Future Large Galaxy Redshift Surveys". In: *Astrophys. J.* 598.2, pp. 720–740. ISSN: 0004-637X. DOI: 10.1086/379122. arXiv: 0307460 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/2003ApJ...598..720S/abstract.

Silva, Marta B. et al. (Jan. 2013). "Intensity Mapping of Lyα emission during the epoch of reionization". In: *Astrophys. J.* 763.2, p. 132. ISSN: 0004-637X. DOI: 10.1088/0004-637X/763/2/132. URL: http://stacks.iop.org/0004-637X/763/i=2/a=132?key=crossref.764db8a2ff3e4e5b9a3497542185a835.

Sinha, Manodeep and Lehman H Garrison (Jan. 2020). "corrfunc – a suite of blazing fast correlation functions on the CPU". In: *Mon. Not. R. Astron. Soc.* 491.2, pp. 3022–3041. ISSN: 0035-8711. DOI: 10.1093/mnras/stz3157. URL: https://academic.oup.com/mnras/article/491/2/3022/5625082.

Slipher, V. M. (1913). "The radial velocity of the Andromeda Nebula". In: *Lowell Obs. Bull.* 1, pp. 56–57. URL: https://ui.adsabs.harvard.edu/abs/1913LowOB...2..56S/abstract.

Slosar, Aňze et al. (2011). "The Lyman-α forest in three dimensions: Measurements of large scale UX correlations from BOSS 1st-year data". In: *J. Cosmol. Astropart. Phys.* 2011.9. ISSN: 14757516. DOI: 10.1088/1475-7516/2011/09/001. arXiv: 1104.5244.

Smee, Stephen A. et al. (2013). "The multi-object, fiber-fed spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey". In:

Astron. J. 146.2. ISSN: 00046256. DOI: 10.1088/0004-6256/146/2/32. arXiv: 1208.2233.

Smeenk, Christopher and George Ellis (2017). *Philosophy of Cosmology*. URL: https://plato.stanford.edu/entries/cosmology/ (visited on 07/29/2020).

Smoot, G. F. et al. (Sept. 1992). "Structure in the COBE differential microwave radiometer first-year maps". In: *Astrophys. J.* 396, p. L1. ISSN: 0004-637X. DOI: 10.1086/186504.

Songaila, Antoinette and Lennox L. Cowie (July 2010). "The Evolution of Lyman Limit Absorption Systems to Redshift Six". In: *Astrophys. J.* 721.2, pp. 1448–1466. DOI: 10.1088/0004-637X/721/2/1448. arXiv: 1007.3262. URL: http://arxiv.org/abs/1007.3262http://dx.doi.org/10.1088/0004-637X/721/2/1448.

Spergel, D. N. et al. (Sept. 2003). " First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters ". In: *Astrophys. J. Suppl. Ser.* 148.1, pp. 175–194. ISSN: 0067-0049. DOI: 10.1086/377226. arXiv: 0302209 [astro-ph]. URL: https://ui.adsabs.harvard.edu/abs/2003ApJS..148..175S/abstract.

Springel, Volker (Dec. 2005). "The cosmological simulation code gadget-2". In: *Mon. Not. R. Astron. Soc.* 364.4, pp. 1105–1134. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2005.09655.x. URL: https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2005.09655.x.

Springel, Volker et al. (Mar. 2018). "First results from the IllustrisTNG simulations: Matter and galaxy clustering". In: *Mon. Not. R. Astron. Soc.* 475.1, pp. 676–698. ISSN: 13652966. DOI: 10.1093/mnras/stx3304. arXiv: 1707.03397. URL: https://ui.adsabs.harvard.edu/abs/2018MNRAS.475..676S/abstract.

Steidel, Charles C. et al. (Aug. 2011). "Diffuse Lyα Emitting Halos: A Generic Property of High-redshift Star-forming Galaxies". In: *Astrophys. J.* 736.2, 160, p. 160. DOI: 10.1088/0004-637X/736/2/160. arXiv: 1101.2204 [astro-ph.CO].

Szapudi, István et al. (Feb. 2001). "Fast Cosmic Microwave Background Analyses via Correlation Functions". In: *Astrophys. J.* 548.2, pp. L115–L118. ISSN: 0004637X. DOI: 10.1086/319105. URL: https://iopscience.iop.org/article/10.1086/319105https://iopscience.iop.org/article/10.1086/319105/meta.

Taniguchi, Yoshiaki, Yasuhiro Shioya, and Yuko Kakazu (Nov. 2001). "On the Origin of Lyα Blobs at High Redshift: Submillimetric Evidence for a Hyperwind Galaxy at z = 3.1". In: *Astrophys. J.* 562.1, pp. L15–L17. ISSN: 0004637X. DOI: 10.1086/338101. URL: http://stacks.iop.org/1538-4357/562/i=1/a=L15.

Viel, Matteo (Oct. 2003). "Cosmology with the lyman-alpha forest in the WMAP era". In: arXiv: 0310413 [astro-ph]. URL: http://arxiv.org/abs/astro-ph/0310413.

Viel, Matteo, Martin G. Haehnelt, and Volker Springel (Nov. 2004). "Inferring the dark matter power spectrum from the Lyman α forest in high-resolution QSO absorption spectra". In: *Mon. Not. R. Astron. Soc.* 354.3, pp. 684–694. ISSN: 00358711. DOI: 10.1111/j.1365-2966.2004.08224.x. URL: https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2004.08224.x.

Vogelsberger, Mark et al. (Jan. 2020). "Cosmological simulations of galaxy formation". In: *Nat. Rev. Phys.* 2.1, pp. 42–66. ISSN: 25225820. DOI: 10.1038/s42254-019-0127-2. arXiv: 1909.07976.

Wadekar, Digvijay et al. (July 2020). "HInet: Generating neutral hydrogen from dark matter with neural networks". In: arXiv: 2007.10340. URL: http://arxiv.org/abs/2007.10340.

Wang, Liang et al. (2015). "NIHAO project - I. Reproducing the inefficiency of galaxy formation across cosmic time with a large sample of cosmological hydrodynamical simulations". In: *Mon. Not. R. Astron. Soc.* 454.1, pp. 83–94. ISSN: 13652966.

DOI: 10.1093/mnras/stv1937. arXiv: 1503.04818. URL: https://ui.adsabs.harvard.edu/abs/2015MNRAS.454...83W/abstract.

Weinberg, David H. et al. (July 1997). "Hubble Flow Broadening of the Lyman-alpha Forest and its Implications". In: *Struct. Evol. Intergalactic Mediu. from QSO Absorpt. Line Syst. Proc. 13th IAP Astrophys. Colloq.* Paris: Editions Frontieres, p. 133. arXiv: 9709303 [astro-ph]. URL: http://arxiv.org/abs/astro-ph/9709303.

Wetzel, Andrew R. et al. (Aug. 2016). "Reconciling dwarf galaxies with ΛCDM cosmology: simulating a realistic population of satellites around a Milky-Way galaxy". In: *Astrophys. J.* 827.2, p. L23. ISSN: 2041-8205. DOI: 10.3847/2041-8205/827/2/123. URL: http://www.tapir.caltech.edu/{~}phopkins/Site/GIZMO.

Wisotzki, L. et al. (2018). "Nearly all the sky is covered by Lyman-α emission around high-redshift galaxies". In: *Nature* 562.7726, pp. 229–232. ISSN: 14764687. DOI: 10.1038/s41586-018-0564-6. arXiv: 1810.00843.

Wolfe, Arthur M., Eric Gawiser, and Jason X. Prochaska (Sept. 2005). "Damped Lyman alpha Systems". In: *Annu. Rev. Astron. Astrophys.* 43, pp. 861–918. DOI: 10.1146/annurev.astro.42.053102.133950. arXiv: 0509481 [astro-ph]. URL: http://arxiv.org/abs/astro-ph/0509481http://dx.doi.org/10.1146/annurev.astro.42.053102.133950.

Yèche, Christophe et al. (June 2017). "Constraints on neutrino masses from Lyman-alpha forest power spectrum with BOSS and XQ-100". In: *J. Cosmol. Astropart. Phys.* 2017.6, p. 047. ISSN: 14757516. DOI: 10.1088/1475-7516/2017/06/047. arXiv: 1702.03314. URL: https://ui.adsabs.harvard.edu/abs/2017JCAP...06..047Y/abstract.

Yue, B. et al. (July 2015). "Intensity mapping of [C II] emission from early galaxies". In: *Mon. Not. R. Astron. Soc.* 450.4, pp. 3829–3839. ISSN: 1365-2966. DOI: 10.1093/mnras/stv933. URL: http://academic.oup.com/mnras/article/450/4/3829/1750783/Intensity-mapping-of-Cii-emission-from-early.