




**Analysis validation of the Lyman- $\alpha$  forest measurements  
from the Dark Energy Spectroscopic Instrument**

**ADVERTIMENT.** L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

**ADVERTENCIA.** El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

**WARNING.** The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

UNIVERSITAT AUTÒNOMA DE BARCELONA  
INSTITUT DE FÍSICA D'ALTES ENERGIES



**Analysis validation of the Lyman- $\alpha$  forest measurements  
from the Dark Energy Spectroscopic Instrument**

Doctorate in Physics  
Programa de Doctorat en Física

César Ramírez Pérez

Thesis supervised by:  
Andreu Font Ribera (Director) and Eduard Massó i Soler (Tutor)

A thesis submitted for the degree of Philosophiae Doctor (PhD)

2024



# ACKNOWLEDGEMENTS

The first person I must thank is Andreu Font-Ribera for all the advice, guidance, and support he has provided me over these years. It has truly been a pleasure to work with such a great person, both professionally and personally. I also want to thank David Alonso for all the discussions and assistance that made this thesis possible; and Ignasi Pérez-Ràfols for our collaboration throughout the year we worked together, and for being a great colleague.

I would like to thank Laura Cabayol, Laura Casas, Andrei Cuceu, James Farr, Calum Gordon, Julien Guy, Hiram Herrera, Jim Rich, Javier Sánchez, the observational cosmology group at IFAE, and many others for their assistance, feedback and contributions that made my research possible.

Special thanks to Laura Cabayol and Umurbt Demirbozan for the interesting conversations; Laura Casas for her uniquely shy boldness; Calum Gordon for his dark humor that resonates so well with mine; Andrea Muñoz for the amusing moments; and the entire Oxford Cosmology group for their warm welcome.

Agradezco a Alejandro todas las charlas y risas a la hora de comer, y a la UAB por proporcionar césped de sobra para ello. A mi madre por preocuparse siempre por los suyos; a mi hermana por cuidar de mí; a mi padre por su apoyo; y a mi suegra por hacerme sentir de su familia y por los muchos momentos de carcajadas incontrolables.

Por último, quiero agradecer a Alicia el haber estado siempre a mi lado estos años. Por entenderme, por levantarme el ánimo, por hacerme feliz, por los buenos momentos en tantos sitios y por todo lo que está por venir.





”

*“Were the succession of stars endless, then the background of the sky would present us a uniform luminosity, like that displayed by the Galaxy — since there could be absolutely no point, in all that background, at which would not exist a star. The only mode, therefore, in which, under such a state of affairs, we could comprehend the voids which our telescopes find in innumerable directions, would be by supposing the distance of the invisible background so immense that no ray from it has yet been able to reach us at all.”*

— **Edgar Allan Poe**, *Eureka* (1848)



# ABSTRACT

In recent decades, advances in galaxy surveys have significantly enhanced our ability to study the universe’s composition and evolution, particularly the elusive dark energy. While galaxy surveys have traditionally been the usual way of studying the distribution of matter, in this thesis I have made important contributions to the development of an alternative probe: the Lyman- $\alpha$  forest.

The Lyman- $\alpha$  forest is a pattern of absorption features in the spectra of distant quasars caused by intervening neutral hydrogen. The Dark Energy Spectroscopic Instrument (DESI) survey will measure the spectra of up to 3 million quasars, providing the best measurement of the universe’s expansion to date using the Lyman- $\alpha$  forest.

My contributions to this survey include the creation of the first Lyman- $\alpha$  catalog using the Early Data Release from the survey (Chapter 6). And the analysis validation using data from Data Release 1 (Chapter 7).

To aid in the development of cosmological surveys, synthetic datasets are used to help build analysis pipelines, estimate covariance matrices, and perform forecasting. In Chapter 5, I present CoLoRe, a code for synthetic data generation for multiple tracers, including the Lyman- $\alpha$  forest. This code allows for the generation of random data from multiple state-of-the-art surveys in the same realisation, enabling the study of the interplay between them.



## PREFACE

The quest to understand the universe's origin, evolution, and fate has captivated human curiosity for centuries. The study of the composition and dynamics of the universe has undergone remarkable advancements, evolving from ancient philosophical speculations to modern scientific inquiries.

Throughout history, humanity's perception of the cosmos has been intertwined with religious, philosophical, and existential concerns. Looking up to the sky and searching for answers amidst the celestial expanse has evoked awe and reverence, fueling contemplation about our place in the universe. From the ancient Greeks' belief in a universe centred around Earth to the inspiring transition to a heliocentric model, cosmology has both shaped and been shaped by cultural, spiritual, and existential narratives. These transitions not only revolutionized our understanding of the universe but also challenged deeply rooted religious and existential convictions.

The 19<sup>th</sup> century witnessed the first successful measurement of the parallax of a few stars outside the Solar System, marking a significant milestone in astronomical observation (Bessel, 1838). However, it was the 20<sup>th</sup> century that truly defined our current understanding of the universe.

The modeling of Cepheid variable stars (Leavitt, 1908) allowed for the measurement of distances to very distant objects. Using this tool, Edwin Hubble demonstrated that Andromeda is far outside the Milky Way (Hubble, 1925), ending the debate over the existence of galaxies beyond our own. This discovery forever changed our conception of the universe, leading to the formalization of the cosmological principle: the universe, on large scales, is isotropic and homogeneous.

Another ancient conception of the universe that had been preserved through the centuries was its static nature. The notion of celestial perfection and unchanging natural laws reinforced this concept, largely due to the lack of observational

evidence to the contrary. For this reason, when Albert Einstein introduced General Relativity (GR) (Einstein, 1915), he added the cosmological constant parameter ( $\Lambda$ ) to the Einstein equations to fulfill the static universe requirement. Later, it was shown by Alexander Friedmann that these equations remain valid in a dynamic universe, regardless of the value of  $\Lambda$ .

The expansion of the universe was observed for the first time by Hubble (Hubble, 1929), leading to the proposal of an expanding universe with a zero cosmological constant (Einstein and de Sitter, 1932). However, by the end of the 20<sup>th</sup> century, measurements of distant supernovae showed that the universe's expansion is accelerating (Riess et al., 1998; Perlmutter et al., 1999), which we now model as a strictly positive cosmological constant. This constant is thought to be caused by a mysterious Dark Energy (DE), which exerts a repulsive force that counteracts the gravitational collapse produced by matter.

During the first two decades of the 21<sup>st</sup> century, cosmology has transformed into its precision era, characterized by datasets of unprecedented size and quality. The Cosmic Microwave Background (CMB) has allowed for studies of the universe at its early stages, while measurements from galaxy surveys capture the statistical distribution of matter in later times. Of particular interest is the use of Baryon Acoustic Oscillations (BAO), which, acting as a standard ruler, has facilitated studies of the expansion and evolution of the universe.

Different techniques are used to map the large scale structure (LSS) of the universe, with two of the most common being spectroscopic and photometric surveys. Photometric surveys measure the brightness of galaxies and quasars in different wavelength bands, which can be used to estimate their approximate distances and types. These surveys also measure shapes, often affected by gravitational lensing. Spectroscopic surveys, on the other hand, obtain optical spectra for a comparatively smaller set of objects, allowing for precise measurements of their distances and velocities. This precision also enables the measurement of Lyman- $\alpha$  absorption caused by neutral hydrogen in the Intergalactic Medium (IGM), observed in the spectra of distant quasars.

In the context of these cosmological surveys, the use of synthetic data has become increasingly necessary to understand and study physical phenomena under controlled conditions. With the use of synthetic data, it is possible to test theories and validate methodologies without the limitations and uncertainties present in real data.

Following the steps of the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al., 2013) and its extension, the Extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al., 2016), the Dark Energy Spectroscopic Instrument

(DESI; DESI Collaboration et al., 2016a) survey will measure the redshifts of about 35 million galaxies and quasars over 14 000 square degrees, covering a redshift range of  $0 < z < 3.5$ . The collaboration is undergoing a five-year campaign to obtain close to a million quasar spectra with  $z > 2$ , which are being used to perform BAO analysis using the Lyman- $\alpha$  forest.

The central core of this doctoral thesis has been carried out in the context of Lyman- $\alpha$  analyses for the DESI Collaboration. In Chapter 6, the generation of the Lyman- $\alpha$  fluctuations catalog for the DESI Early Data Release (EDR) is described, which was published in Ramírez-Pérez et al. (2024). In Chapter 7, part of the validation of the Lyman- $\alpha$  BAO analysis for the Data Release 1 (DR1) of DESI is described, this was my contribution to the article published in DESI Collaboration et al. (2024b).

I have also worked on the generation of synthetic data, including Lyman- $\alpha$  forest but also lensing and galaxy clustering. In Chapter 5, the presentation and validation of the software used are discussed. This Chapter follows the publication Ramírez-Pérez et al. (2022).





# CONTENTS

<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xxv</b>
<b>Acronyms</b>	<b>xxvii</b>
<b>1 The homogeneous universe</b>	<b>1</b>
1.1 General Relativity for cosmology . . . . .	1
1.1.1 Curved spacetimes . . . . .	3
1.1.2 The Einstein Equations . . . . .	3
1.2 The expanding universe . . . . .	4
1.2.1 Distances in the expanding universe . . . . .	4
1.2.2 Redshift . . . . .	5
1.2.3 The Hubble law . . . . .	6
1.3 The Friedman-Lamaître-Robertson-Walker metric . . . . .	7
1.4 Cosmic ingredients . . . . .	8
1.5 Thermal history . . . . .	10
<b>2 Origin and evolution of inhomogeneities</b>	<b>13</b>
2.1 Inflation . . . . .	13
2.2 Linear evolution . . . . .	15
2.2.1 Evolution of the perturbations . . . . .	15
2.2.2 Evolution of species . . . . .	16
2.2.3 Photon-baryon oscillations . . . . .	17
2.2.4 The two point correlation function . . . . .	18
2.2.5 Non-linear evolution . . . . .	20
<b>3 Cosmological Probes</b>	<b>23</b>

3.1	The Cosmic Microwave Background . . . . .	23
3.2	Galaxy clustering . . . . .	24
3.2.1	Clustering in redshift space . . . . .	25
3.2.2	BAO measurements from galaxy clustering . . . . .	28
3.3	The Lyman- $\alpha$ forest . . . . .	28
3.4	Weak Gravitational Lensing . . . . .	32
<b>4</b>	<b>The Dark Energy Spectroscopic Instrument</b>	<b>35</b>
4.1	Spectroscopic pipeline . . . . .	36
4.2	Data Releases . . . . .	37
<b>5</b>	<b>Generating synthetic datasets for cosmological probes with CoLoRe</b>	<b>39</b>
5.1	Introduction . . . . .	40
5.2	Methods . . . . .	41
5.2.1	Overall code structure . . . . .	41
5.2.2	Cosmological assumptions . . . . .	44
5.2.3	Matter box . . . . .	45
5.2.4	Tracers . . . . .	48
5.2.5	Lognormal predictions . . . . .	58
5.3	Results . . . . .	59
5.3.1	Validation . . . . .	59
5.3.2	Performance at scale . . . . .	66
5.4	Conclusions . . . . .	72
5.5	Appendix: Higher-order terms in the modelling of redshift-space distortions . . . . .	74
<b>6</b>	<b>The Lyman-<math>\alpha</math> forest catalog from the Dark Energy Spectroscopic Instrument Early Data Release</b>	<b>79</b>
6.1	Introduction . . . . .	80
6.2	Data . . . . .	81
6.2.1	DESI spectroscopic data . . . . .	82
6.2.2	Description of the quasar catalog . . . . .	84
6.2.3	BAL and DLA information . . . . .	85
6.3	Spectral reduction . . . . .	85
6.3.1	Masks (sky lines, galactic absorption, DLA, BAL) . . . . .	87
6.3.2	Re-calibration . . . . .	90
6.3.3	Continuum fitting . . . . .	95
6.3.4	Changes with respect to previous analyses . . . . .	100
6.4	Discussion . . . . .	102

6.4.1	Optimal weights . . . . .	103
6.4.2	Selection of rest-frame wavelength range . . . . .	105
6.4.3	Per-quasar parameters (a,b) . . . . .	107
6.5	Summary and conclusions . . . . .	109
<b>7</b>	<b>Robustness tests for the DESI DR1 Lyman-<math>\alpha</math> forest BAO analysis</b>	<b>111</b>
7.1	Data splits . . . . .	113
7.2	Alternative analyses . . . . .	117
7.2.1	Variations in the estimation of fluctuations . . . . .	117
7.2.2	Variations in the measurement of correlations . . . . .	119
7.2.3	Variations in the estimation of cosmological and astrophysical parameters . . . . .	120
<b>8</b>	<b>Conclusion</b>	<b>123</b>
	<b>Bibliography</b>	<b>125</b>



# LIST OF FIGURES

1.1	Energy content of the universe as a function of the scale factor $a$ . The shaded regions mark the species that dominate at each stage of the expansion history. At early times, radiation was dominant; at $z \sim 3900$ , the matter-radiation equality occurred; and from then on, matter was dominant. Very recently, at $z \sim 0.30$ , DE began to dominate. . . . .	10
2.1	The (linear theory) matter power spectrum fitted from multiple cosmological probes. The data markers correspond to inferences at $z = 0$ based on different clustering measurements. . . . .	19
3.1	Top: Lyman- $\alpha$ auto-correlation along with the best fit model from the DESI DR1 BAO analysis. The different colors correspond to different orientations with respect to the line-of-sight, with blue correlations being close to the line-of-sight $0.95 < \mu < 1$ . Bottom: Measured Lyman- $\alpha$ cross-correlation with quasars for the same survey. (DESI Collaboration et al., 2024b) . . . . .	31
3.2	Effects of shear ( $\kappa$ ) and convergence (with components $\gamma_{1,2}$ ) in the shape of an originally circular galaxy. The convergence magnifies the image, while the two shear components stretch the image in different directions (Bovy, 2024). . . . .	33
5.1	Bias models implemented in CoLoRe. The exponential model preserves the “lognormality” of the field if using the lognormal structure formation model, but it can lead to numerically unstable results in the presence of sufficiently large fluctuations in the Gaussian field. The exp-truncated model can be used to curb this behaviour. . . . .	49

5.2	Simulated maps of the anisotropic stochastic gravitational wave background from astrophysical sources at redshifts $z < 0.4$ using the models of (Cusin et al., 2020). The top and bottom plots show simulations using the lognormal and first-order Lagrangian Perturbation Theory (LPT) structure formation models respectively. The former is characterised by strong positive fluctuations on a few regions, while the latter displays the more physical filamentary structure of the cosmic web. . . . .	50
5.3	<i>Top</i> : simulated map of the low-redshift ISW effect for a source plane at $z_* = 0.5$ . <i>Bottom</i> : map of the lensing convergence for a source plane at $z_* = 1.4$ . . . . .	52
5.4	Maps of source-related quantities for a simulated CoLoRe catalog in the redshift range $z < 0.3$ . <i>Top left</i> : source overdensity. <i>Top right</i> : mean redshift distortion. <i>Middle</i> : mean lensing shear. <i>Bottom</i> : mean lensing displacement vector. . . . .	54
5.5	Density ( <i>top</i> ) and radial velocity ( <i>bottom</i> ) skewers for two arbitrary sources in a simulated CoLoRe catalog. . . . .	54
5.6	Slice through one of the 21cm intensity maps generated with CoLoRe. . . . .	57
5.7	Redshift distribution of the two tomographic bins used in the analysis of the validation simulations (red and blue lines), as well as the overall redshift distribution (black dashed line). The bins are defined by a cut in photometric redshift space at $z_{\text{photo}} = 0.7$ , where we assigned each source a random photometric redshift error with standard deviation $\sigma_z = 0.03(1 + z)$ . . . . .	60
5.8	Galaxy clustering, cosmic shear and lensing displacement power spectra. The gray bands show the 68% scatter from the 100 validation realizations, and the theoretical predictions, described in Section 5.2.5, are shown as black solid lines. The lower-left corner shows all auto- and cross-correlations between the two galaxy clustering and cosmic shear bins. The upper-right corner shows the auto- and cross-correlations between the lensing displacement vectors in both redshift bins. . . .	61
5.9	Cross-correlations between ISW (top row) and convergence maps (bottom row) at $z = 1$ and the two high-redshift clustering and shear samples in the validation simulations. The gray bands show the 68% scatter from the 100 validation realizations, and the theoretical predictions, described in Section 5.2.5, are shown as black solid lines. . . . .	62

5.10	Shear power spectra for the first redshift bin in the validation simulations. The blue and red points show the results (mean and standard deviation of the 100 simulations) with no correction for the effects of source clustering. The gray points show the result of applying this correction by simply subtracting the <i>B</i> -mode power spectrum from the <i>E</i> -mode one. The black line shows the theoretical prediction for the latter. The orange band shows the $1\sigma$ uncertainties one would find in the presence of realistic shape noise, which would make the source clustering effect undetectable in practice. . . . .	62
5.11	Measurements of the correlation function from the stack of 10 realizations used to validate the 3D clustering. The low- <i>z</i> samples take redshifts from 0.5 to 0.7, while high- <i>z</i> samples take redshifts from 0.7 to 0.9. The lines show the model, solid in the regions where the bias was fitted, and the shaded bands show the error for a single realization. <i>Top</i> : Measurements of the monopole in real space. <i>Bottom</i> : Measurements of the monopole and quadrupole in redshift space. . . . .	65
5.12	Redshift distribution of the different tracers simulated in our two flagship simulations. These include the LSST gold sample (black), the 4 DESI samples (blue) and 4 different radio continuum samples observable by SKA (red). . . . .	66
5.13	Visual description of the multi-tracer products that can be simulated with CoLoRe. The upper plot shows one of the beams used internally by CoLoRe for domain decomposition, with the redshift and angular coordinates of 3 different DESI samples, and maps of the density, radial velocity, and lensing shear constructed from sources at two different redshifts. The lower plot shows the density skewers calculated for three arbitrary DESI quasars contained in the same beam, as a function of comoving distance. . . . .	68



5.14	Fraction of the total run time taken up by different stages of a typical CoLoRe simulation. The stages shown are, the generation of the initial Gaussian random fields in Fourier space ("GRF"), their transformation to real space ("FFT"), the structure formation model leading to a positive-definite matter overdensity ("LN" or "1LPT" for the lognormal and first-order LPT simulations), the generation of source catalogs via Poisson sampling ("Poisson"), the redistribution of these sources across different nodes before any line-of-sight calculations ("Source redistrib."), the calculation of all relevant line-of-sight quantities ("line of sight tracing"), and the output of all final products to disk ("Write"). Note that the line of sight tracing stage is more sensitive to inter-node communication than other stages, and thus it takes up the same fraction of the total compute time in both simulations in spite of the additional time taken by the 1LPT stage, given the larger number of MPI nodes needed for that simulation. . . . .	69
5.15	Relative difference between the shear power spectrum of a CoLoRe simulation run using the "fast lensing" scheme, and a simulation run without this approximation. The effects caused by the various interpolations carried out as part of the fast scheme should be carefully modelled if an accurate theoretical description of the CoLoRe output is required. .	71
5.16	Measurements and predictions for the different cross-correlations discussed in Section 5.5 for a single simulation with two special tracers ( $b_A = 0.001$ and $b_B = 2$ ). The predictions are from the simpler model discussed in Section 5.3.1.3 and are missing those terms involving $\epsilon(\mathbf{x}) = \delta_{\text{LN}}(\mathbf{x})\eta(\mathbf{x})$ . This explains the disagreement seen on the small-scales quadrupole in panel (c), it also shows a small disagreement in panels (e) and (f). . . . .	75
5.17	Contributions to the monopole (top panel/blue lines and bands) and the quadrupole (bottom panel/red lines and bands) from the terms in Eqs. Eq. (5.42)-Eq. (5.44) using the same simulation as in Figure Fig. 5.16. The shaded bands show the error in the measurement of each term from the scatter between 48 healpixels of $N_{\text{side}} = 2$ . The dashed lines show the full model prediction as in Section 5.3.1.3, where the extra terms are not included. Particularly the term $\langle \epsilon \epsilon \rangle$ (right panel) has an important impact on the small-scales quadrupole. . . . .	76

6.1	Distribution of wavelength and pipeline-reported error on the flux measurement for the larger EDR+M2 sample. This measurement is performed in the C III region of the spectra, as defined in Table 6.2. The solid lines mark the mean value of the error for a given wavelength, black for the EDR sample and white for the EDR+M2. Apart from the clear distinction in these two lines, a subtle division into two bands at larger wavelengths can also be observed. This is caused by the better signal-to-noise in the EDR sample, thanks to multiple re-observations of the same targets. We can also observe how the variance relatively increases in the two ends of the spectrograph and in the area affected by the collimator mirror reflectivity around 4400 Å. . . . .	83
6.2	Distribution of objects in the two samples used in this publication, compared with the final eBOSS DR16 sample presented in (du Mas des Bourboux et al., 2020). Top: Redshift distribution of the three catalogs. Bottom: Sky distribution of the same catalogs. We observe that M2 makes a significant portion of the EDR+M2 sample. When compared to the eBOSS data, the EDR+M2 sample is approximately halfway to matching the number of objects in the former. . . . .	86
6.3	Fraction of pixels masked due to Damped Lyman- $\alpha$ Absorption (DLA)s (top) and Broad Absorption Line (BAL) (bottom) features. As expected, the number of detected DLAs increase with redshift (and therefore with $\lambda$ ), yielding a fraction of masked pixels always below the 5%. For the BAL case, we show the masked fraction as a function of $\lambda_{\text{RF}}$ , which allow us to observe the strong wavelength dependence of the masking, associated with emission lines for different elements. This Figure used the full EDR+M2 dataset. . . . .	89
6.4	Weighted average of the flux-transmission field measured in the C III region. This measurement has been performed without masking sharp features, and they can be clearly observed at different positions in the spectrograph. The three masked regions are specified in the plot, showing the wavelength range affected by the mask. Smooth features in this measurement can also be observed. Their impact can be corrected through the flux re-calibration (see Section 6.3.2). For this Figure we used the full EDR+M2 dataset. . . . .	91

- 6.5 Bottom: Weighted average of the flux-transmission field measured at different regions of the spectra. Its value has been shifted to better distinguish the different regions. As opposed to the results shown in Fig. 6.4, sharp features are not present here because these samples have already been masked. The smooth features in the spectra are similar for all the measured regions, being the Mg II-R an outlier in this tendency, likely to be caused due to the reduced number of pixels available for this region at low wavelengths (see Fig. 6.6). Results are computed from the full EDR+M2 sample. Top: Average residuals to White Dwarf (WD) spectra in the blue arm of the DESI spectra, as seen in Guy et al. (2023). A similar trend can also be observed here between these residuals and our measured average of the flux-transmission field, especially at smaller wavelengths, further justifying our re-calibration process. . . . . 93
- 6.6 Number of pixels available for the different regions measured in bins of  $\Delta\lambda = 55.58 \text{ \AA}$ . Given the location of the different regions at different positions in the quasar spectra, the number of available pixels is different for each of them. This is because our spectral coverage lies in the range  $(3600, 5772) \text{ \AA}$ . Thanks to the larger number of pixels available for the C III region, it was selected as the region to be used for the re-calibration process. This Figure used the full EDR+M2 sample. . . 94
- 6.7 Wavelength evolution of the fitted parameters  $\eta$  and  $\sigma_{\text{LSS}}^2$ , measured in both the C III and Lyman- $\alpha$  regions for the EDR (dashed) and EDR+M2 (solid) samples. Top: The pipeline error correction  $\eta$  is found to be larger for the EDR sample, caused by the worse estimation of the pipeline-reported variance for the first part of the SV program. Bottom:  $\sigma_{\text{LSS}}^2$ , in this case it is consistent between the two samples. As expected the C III region shows a value close to 0 for all wavelengths, while for the Lyman- $\alpha$  region follows the expected increase in its intrinsic variance with redshift. In both  $\eta$  and  $\sigma_{\text{LSS}}^2$  measurements for the EDR sample, the fitted parameters could not be obtained for the larger wavelength value due to the reduced number of pixels available, falling to the default values  $\eta = 0.1$  and  $\sigma_{\text{LSS}}^2 = 0.1$ . . . . . 98

6.8	Measurement of the weighted mean of the flux-transmission field for both the Lyman- $\alpha$ and C III regions. The Lyman- $\alpha$ region shows an expected higher variance at all wavelengths compared to the C III region, although it lacks smooth features thanks to the re-calibration process. Similarly, the EDR sample shows a higher variance than EDR+M2. In both cases, this is caused by the decrease in number of pixels available. Given the low number of pixels available at larger wavelengths for the EDR sample for the Lyman- $\alpha$ region, it departs from the expected unity behavior. . . . .	99
6.9	Example of a high signal-to-noise quasar spectrum. The mean expected flux $\overline{C}(\lambda_{\text{RF}})$ as in Eq. (6.4) is shown for the Lyman- $\alpha$ , S IV, C IV and C III regions. The quasar has a redshift $z_q = 2.495$ and is identified as a DESI object with TARGETID=39628443918272474. As in this example, we occasionally observe metal absorption in the calibration regions. .	100
6.10	Contribution of each term in Eq. (6.9) to the full variance, for multiple wavelength bins. For this measurement we only used the 5% of quasars with the highest SNR in the EDR+M2 sample. This analysis reveals that the effect of the $\epsilon$ term in Eq. (6.9) is minimal, even for the data subset where it is supposed to be most significant. . . . .	102
6.11	Measurement of the Lyman- $\alpha$ auto- (solid) and cross- (dashed) correlation errorbars for different choices of the $\sigma_{\text{mod}}^2$ parameter, scaled to a reference value at $\sigma_{\text{mod}}^2 = 1$ . The value of the errorbars was averaged over all scales, given the small scale dependency. The $\sigma_{\text{mod}}^2$ parameter modified the inverse-variance weighting scheme as defined in Eq. (6.10), the use of a $\sigma_{\text{mod}}^2 \neq 1$ allows us for the optimization of the weighting scheme. In both auto- and cross-correlations, we observe a reduction in the size of the errorbars when we approach the optimal value of the $\sigma_{\text{mod}}^2$ parameter. This optimal value is slightly different, but in both cases is found around 7-8. In the optimal value, the improvement in the measurement of the auto-correlation is about 20%, and 10% for the case of the cross-correlation. Both measurements have been performed with the full EDR+M2 dataset. . . . .	104

- 6.12 Comparison of the size of the errorbars in the Lyman- $\alpha$  auto-correlation for different choices of  $\lambda_{\text{RF}, \text{max}}$ . The blue points show the errorbars size after averaging over all scales (given the small scale dependence of this quantity), and scaled to a reference value at  $\lambda_{\text{RF}, \text{max}} = 1205 \text{ \AA}$ . The orange points show the equivalent measurement but in this case for the product of errorbar sizes times the number of pixel pairs, removing the dependence on the number of pairs used in the measurement. Following the blue points, we observe an improvement in the precision of our auto-correlation measurements when increasing the  $\lambda_{\text{RF}, \text{max}}$  parameter, having the optimal value at  $1205 \text{ \AA}$ . The orange points show that the quality of these added points actually decrease for higher wavelengths, revealing that the improved performance is only driven by the inclusion of more information in our sample. We selected  $1205 \text{ \AA}$  as our default value as a compromise between these two features. We used data from the whole EDR+M2 dataset for these measurements. 106
- 6.13 Identical measurements as the ones displayed in Fig. 6.12, in this case for values of  $\lambda_{\text{RF}, \text{min}}$ . The optimal wavelength is found at  $\lambda_{\text{RF}, \text{min}} = 1040 \text{ \AA}$ , while having similar features for the orange points: a decrease in the quality of the points added when approaching the emission line. Following the same arguments as in the case of  $\lambda_{\text{RF}, \text{max}}$ , we decided to keep the limit at  $1040 \text{ \AA}$ . Again, these measurements were performed using the EDR+M2 sample. . . . . 107
- 6.14 Distribution of  $a_q$  and  $b_q/a_q$  parameters defined in Eq. (6.4) for the EDR+M2 sample. The  $a_q$  parameter modifies the amplitude of the mean continuum  $\overline{C}(\lambda_{\text{RF}})$ , while the  $b_q/a_q$  term introduces a modification that tilts it as a function of rest-frame wavelength. This last parameter relates to the intrinsic width of the spectral index. The faint long tail at large values of  $|b_q/a_q|$  is caused by short forests in the sample, where the continuum fit can be problematic. . . . . 108

7.1	Measurements of the BAO parameters along the line of sight ( $\alpha_{\parallel}$ ) and across the line of sight ( $\alpha_{\perp}$ ) with contours corresponding to the 68% and 95% confidence regions. The auto-correlation results (filled blue contours) are the combined measurement of the Ly $\alpha$ forest auto-correlations in the Lyman- $\alpha$ and Lyman- $\beta$ regions. The cross-correlation results (dashed black) are the correlations of the forest in these two regions with quasars. The combined results (solid red) simultaneously fit all four correlations taking into account their cross-covariance (DESI Collaboration et al., 2024b). . . . .	113
7.2	BAO constraints from the main analysis (grey) and from data splits. Top left: low (green) vs high (blue) SNR in the quasar spectrum. Top right: low (green) vs high(blue) C iv equivalent width (EW) in the quasar spectrum. Bottom left: South (green) vs North (blue) imaging used in the quasar target selection. Bottom right: correlations from the Lyman- $\alpha$ region (green) and Lyman- $\beta$ region (blue); the Lyman- $\alpha$ region shows the combined measurement from the auto-correlations of the forest measured in the Lyman- $\alpha$ region and the cross-correlations of this region with quasars. The contours labeled Lyman- $\beta$ show the combined measurement of the forest auto-correlations measured in the Lyman- $\beta$ region and the cross-correlation of this region with quasars. . . . .	114
7.3	Shifts in the BAO parameters from alternative analyses. These include variations in the method to estimate the fluctuations (purple); variations in the dataset used (red); variations in the measurement of correlations and covariances (green); variations in the range of separations used (orange); and variations in the modelling (blue). The red shaded region shows the one $\sigma$ uncertainty from the main analysis, and the smaller gray region shows the threshold set for these tests ( $\sigma/3$ ). . . . .	116



# LIST OF TABLES

5.1	Simulation products generated by CoLoRe. The first 11 rows show the different tracers generated for the large-volume simulations described in the text. For each tracer we show the physical quantities simulated (Redshift Space Distortions (RSD), lensing displacements, shear, convergence, and density/velocity skewers), as well as the memory and disk space taken. The last two rows display the memory requirements for the different Cartesian grids stored for lognormal and 1LPT simulations. Although the memory requirements are dominated by these cartesian grids (particularly for LPT simulations), the simulated tracers can take up a non-negligible fraction of the available memory. This is patent for the LSST sample, given its size, and the need to store lensing information, and for the DESI quasar sample, since we save a full density/velocity skewer for each source. . . . .	70
6.1	Number of Lyman- $\alpha$ quasars in the two samples, number of quasars showing BAL features and number of DLAs affecting forests from the Lyman- $\alpha$ quasars. The better signal-to-noise in the EDR sample allows for the detection of a higher number of DLA and BAL features. . . .	87
6.2	Statistics for the regions considered during the analysis, the region span is defined in the quasar rest-frame wavelength ( $\lambda_{\text{RF}}$ ). The number of forests corresponds to the number of quasars whose spectra can be observed in the spectrograph. A minor number of forests are rejected due to low signal-to-noise ratio (SNR) or due to them being too short.	92





# ACRONYMS

<b>BAL</b>	Broad Absorption Line ( <i>pp. xix, xxv, 81, 85, 87–89, 114, 115, 118</i> )
<b>BAO</b>	Baryon Acoustic Oscillations ( <i>pp. viii, ix, xv, xxiii, 9, 17, 28, 30–32, 35, 37, 59, 80, 102, 103, 105, 109, 111–114, 116, 117, 120, 121, 123, 124</i> )
<b>BBN</b>	Bick Bang Nucleosynthesis ( <i>pp. 11, 28</i> )
<b>BOSS</b>	Baryon Oscillation Spectroscopic Survey ( <i>pp. viii, 30, 35, 80, 111, 124</i> )
<b>CDM</b>	Cold Dark Matter ( <i>p. 8</i> )
<b>CMB</b>	Cosmic Microwave Background ( <i>pp. viii, 8, 9, 11, 19, 23, 24, 28, 40, 41, 66, 67</i> )
<b>COBE</b>	Cosmic Background Explorer ( <i>p. 24</i> )
<b>DE</b>	Dark Energy ( <i>pp. viii, xv, 4, 9–11, 16, 17, 35, 40, 44</i> )
<b>DES</b>	Dark Energy Survey ( <i>pp. 28, 33</i> )
<b>DESI</b>	Dark Energy Spectroscopic Instrument ( <i>pp. viii, ix, xv, xxi, 25, 28, 30–32, 35–37, 39, 40, 70, 72, 79–85, 87, 88, 90, 91, 94, 95, 100, 102, 105, 109–112, 115, 117, 123, 124</i> )
<b>DLA</b>	Damped Lyman- $\alpha$ Absorption ( <i>pp. xix, xxv, 85, 87–89, 109, 118, 120</i> )
<b>DR1</b>	Data Release 1 ( <i>pp. ix, xv, 28, 30, 31, 37, 82, 111, 112, 118, 124</i> )
<b>eBOSS</b>	Extended Baryon Oscillation Spectroscopic Survey ( <i>pp. viii, xix, 30, 35, 37, 59, 80, 82, 85, 86, 102, 109, 111, 117–120, 124</i> )
<b>EDR</b>	Early Data Release ( <i>pp. ix, xix–xxi, 32, 37, 79–81, 83, 97–99, 102, 109, 111, 118, 124</i> )
<b>EDR+M2</b>	Early Data Release plus the first two months of the main survey ( <i>pp. xix–xxii, 81–83, 85–87, 89, 91–94, 97–99, 102, 104, 106–109</i> )
<b>FLRW</b>	Friedman-Lamaître-Robertson-Walker ( <i>pp. 7, 13, 15</i> )

<b>GR</b>	General Relativity ( <i>pp.</i> <a href="#">1</a> , <a href="#">2</a> , <a href="#">6</a> , <a href="#">32</a> )
<b>IGM</b>	Intergalactic Medium ( <i>pp.</i> <a href="#">viii</a> , <a href="#">11</a> , <a href="#">28</a> , <a href="#">89</a> , <a href="#">109</a> )
<b>ISM</b>	Interstellar Medium ( <i>p.</i> <a href="#">90</a> )
<b>LPT</b>	Lagrangian Perturbation Theory ( <i>pp.</i> <a href="#">xvi</a> , <a href="#">xviii</a> , <a href="#">xxv</a> , <a href="#">20</a> , <a href="#">21</a> , <a href="#">27</a> , <a href="#">40</a> , <a href="#">41</a> , <a href="#">47</a> , <a href="#">48</a> , <a href="#">50</a> , <a href="#">51</a> , <a href="#">56</a> , <a href="#">67</a> , <a href="#">69</a> , <a href="#">70</a> , <a href="#">72</a> , <a href="#">73</a> , <a href="#">123</a> )
<b>LSS</b>	large scale structure ( <i>pp.</i> <a href="#">viii</a> , <a href="#">32</a> , <a href="#">51</a> , <a href="#">58</a> )
<b>LSST</b>	Legacy Survey of Space and Time ( <i>pp.</i> <a href="#">25</a> , <a href="#">39</a> , <a href="#">40</a> )
<b>RSD</b>	Redshift Space Distortions ( <i>pp.</i> <a href="#">xxv</a> , <a href="#">26</a> , <a href="#">27</a> , <a href="#">65</a> , <a href="#">70</a> , <a href="#">73</a> )
<b>SDSS</b>	Sloan Digital Sky Survey ( <i>pp.</i> <a href="#">28</a> , <a href="#">81</a> , <a href="#">87</a> , <a href="#">100</a> , <a href="#">101</a> )
<b>SKA</b>	Square Kilometer Array ( <i>pp.</i> <a href="#">39</a> , <a href="#">40</a> )
<b>SNR</b>	Signal-to-noise ratio ( <i>pp.</i> <a href="#">113</a> , <a href="#">114</a> )
<b>SV</b>	Survey Validation ( <i>pp.</i> <a href="#">37</a> , <a href="#">79</a> )
<b>WMAP</b>	Wilkinson Microwave Anisotropy Probe ( <i>p.</i> <a href="#">24</a> )

# THE HOMOGENEOUS UNIVERSE

Cosmology is the branch of science that studies the origin, evolution, structure, and future of the universe. On large scales, the effects can be primarily described by gravity. Therefore, gravity becomes the main tool used to understand our universe.

Our current framework for describing the effects of gravity began in the early 20<sup>th</sup> century with the development of General Relativity (GR) by Albert Einstein (Einstein, 1915). GR closely relates the content of the universe to its dynamics, providing us with a tool to understand the universe as a whole.

According to the cosmological principle, the universe is homogeneous on large scales. A correct description of these scales provides us with a general picture, which will be further enhanced in the following Chapters. In these Chapters, we will explore the inhomogeneities appearing at intermediate to low scales, which are necessary to explain the structure of our universe as observed today.

This Chapter will begin by introducing the background tools from GR that aid in understanding the cosmological concepts discussed in this thesis. Subsequently, we will explore the solutions of the Einstein Equations that better explain the properties of our universe on large scales. Following this, we will analyze the particles present in the universe. Finally, we will provide a brief overview of the methods used to measure distances in the expanding universe.

## 1.1 General Relativity for cosmology

GR allows us to mathematically model the behavior of gravity on large scales. In this framework, the concept of manifolds is of paramount importance. A manifold is a topological space that locally resembles Euclidean space, providing the tools to model the more complex and typically non-linear structures of the cosmos.

The geometric and causal structure of a given spacetime can be captured in the metric tensor  $g_{\mu\nu}$ , which mathematically allows for the definition of distances and angles. In particular, the length element between two events, called the interval, can be defined as:

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \quad (1.1)$$

where  $dx^\mu$  represents the components of an infinitesimal coordinate displacement four-vector, and where we sum over repeated indices following the Einstein convention.

The interval provides information about the causal structure of spacetime. When  $ds^2 < 0$ , the interval is timelike, and can be traversed by a massive object. If  $ds^2 = 0$ , we have a lightlike interval, which can only be traversed by massless objects. Finally, for  $ds^2 > 0$ , we have a spacelike interval, which cannot be traversed because it connects two causally disconnected points in spacetime.

In the case of an empty universe with no massive objects, the geometry of spacetime is completely determined by the 4-dimensional Minkowski metric:

$$g_{\mu\nu} = \eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +1 \end{pmatrix} \quad (1.2)$$

this is the case of Special Relativity, derived by Einstein in 1905 (Einstein, 1905). In the case of GR, the complexity of the metric increases because it will be dependent on the content of the universe and how this content is distributed.

We can now introduce the Equivalence Principle, which lies at the core of the theory. The weak version of the principle states:

*The motion of freely-falling particles are the same in a gravitational field and a uniformly accelerated frame, in small enough regions. (Carroll, 2001)*

essentially meaning that gravitational pull and local acceleration are equivalent. The “strong” and generalized version of the principle states:

*In small enough regions of spacetime, the laws of physics are those of Special Relativity. (Carroll, 2001)*

being the underlying assumption here that the presence of gravity makes Special Relativity inconsistent by affecting the curvature of spacetime.

### 1.1.1 Curved spacetimes

We need a theory of gravitation that fulfills the Equivalence Principle, and therefore we need to work with curved topological spaces that resemble the Euclidean space and are differentiable at each point.

In such manifolds, we need to upgrade our usual derivative to the covariant derivative, which will behave appropriately in curved spacetimes:

$$\nabla_\mu F_\gamma{}^\nu = \partial_\mu F_\gamma{}^\nu + \Gamma_{\sigma\mu}^\nu F_\gamma{}^\sigma - \Gamma_{\gamma\mu}^\sigma F_\sigma{}^\nu, \quad (1.3)$$

where  $\partial_\alpha \equiv \frac{\partial}{\partial x^\alpha}$ ,  $F$  a generic vector field and:

$$\Gamma_{\alpha\beta}^\mu = \frac{g^{\mu\nu}}{2} (\partial_\beta g_{\alpha\nu} + \partial_\alpha g_{\beta\nu} - \partial_\nu g_{\alpha\beta}), \quad (1.4)$$

the Christoffel symbols. These contain the information from the curvature. Although the curvature of a manifold is usually expressed by using the Riemann curvature tensor:

$$R^\alpha{}_{\mu\nu\rho} = \partial_\nu \Gamma_{\mu\rho}^\alpha - \partial_\rho \Gamma_{\mu\nu}^\alpha + \Gamma_{\mu\rho}^\beta \Gamma_{\nu\beta}^\alpha - \Gamma_{\mu\nu}^\beta \Gamma_{\rho\beta}^\alpha. \quad (1.5)$$

In GR, we use contractions of this whole tensor, the Ricci tensor:

$$R_{\mu\nu} = g^{\alpha\lambda} R_{\alpha\mu\lambda\nu}, \quad (1.6)$$

and the Ricci scalar:

$$R = R^\mu{}_\mu. \quad (1.7)$$

### 1.1.2 The Einstein Equations

The set of equations that fulfill the Equivalence Principle are the Einstein equations. They describe the relationship between the content of the universe (through the energy-momentum tensor  $T_{\mu\nu}$ ) and its curvature (through the Ricci scalar and Ricci tensor):

$$G_{\mu\nu} + \Lambda g_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G_N T_{\mu\nu} \quad (1.8)$$

where the coefficient  $8\pi G_N$  ensures that we fulfill the Newtonian limit when describing flat manifolds. The cosmological constant  $\Lambda$  will play an important role when describing the expanding universe in Section 1.2.

The energy-momentum tensor  $T_{\mu\nu}$  is a physical quantity that generalizes the stress tensor of Newtonian physics, describing the density and flux of energy

and momentum. On large scales, while assuming a perfect fluid, this can be parametrized by the rest-frame density  $\rho$  and the isotropic pressure  $P$ :

$$T^\mu_\nu = \text{diag}(\rho, -P, -P, -P). \quad (1.9)$$

The conservation of energy-momentum is then expressed as:

$$\nabla_\mu T^\mu_\nu = 0. \quad (1.10)$$

## 1.2 The expanding universe

When Einstein derived the field equations in Eq. (1.8), he assumed a static universe. To achieve this, he added a term to the equations to counterbalance the effect of gravity. But the discovery in 1928 by Edwin Hubble that the universe is actually expanding and thus not static (Hubble, 1929) forced Einstein to remove  $\Lambda$  from his field equations.

Much later, in 1998, measurements of distant supernovae showed that the expansion of the universe is actually accelerating (Riess et al., 1998; Perlmutter et al., 1999), thereby offering the opportunity to reintroduce the cosmological constant  $\Lambda$ . The current  $\Lambda$ CDM model assumes that the term  $\Lambda$  in the Einstein field equations is the cause of a Dark Energy (DE) that permeates all the space in the universe and leads to its accelerated expansion.

### 1.2.1 Distances in the expanding universe

Cosmological research relies on the ability to observe and measure distances across the universe. However, the task is complicated by the fact that the universe is expanding and could also have intrinsic curvature. Given this complexity, different distance measures are used depending on convenience.

In general, to express the expansion of the universe, the scale factor  $a(t)$  is used. This parameter evolves over time, describing the normalized size of the universe. For two objects that are not gravitationally bound, we have  $a(t)/a(t_0) = s(t)/s(t_0)$ , where  $s(t)$  is the proper distance between the two objects at time  $t$ . We usually set  $a(t_0) = 1$ , normalizing the scale factor to 1 for the present time, decreasing in the past and growing in the future with expansion.

It is useful to define a distance that factors out the expansion of the universe, and this is the comoving distance  $\chi$ . For an object that emitted light at time  $t$ , it will be at a comoving distance of:

$$\chi(t) = \int_t^{t_0} \frac{1}{a(t)} dt. \quad (1.11)$$

We can relate the comoving distance and the proper distance (or physical distance) through the scale factor:

$$\chi(t) = s(t)/a(t). \quad (1.12)$$

Another relevant measurement of distance is the angular diameter distance  $D_A$ , which relates the physical size of an object to its observed angular size. Using the small angle approximation:

$$D_A = \frac{s}{\theta} = S_k(\chi)a(t), \quad (1.13)$$

where  $s$  is the physical size of the object,  $\theta$  is the angular size,  $a(t)$  the scale factor at emission time and  $S_k$  the comoving transverse distance, which depends on the curvature density parameter  $\Omega_k$  (see Eq. (1.29)) through:

$$S_k(\chi) = \begin{cases} \sinh(\chi H_0 \sqrt{\Omega_k}) / (H_0 \sqrt{\Omega_k}), & \Omega_k > 0 \\ \chi, & \Omega_k = 0 \\ \sin(\chi H_0 \sqrt{|\Omega_k|}) / (H_0 \sqrt{|\Omega_k|}), & \Omega_k < 0 \end{cases} \quad (1.14)$$

### 1.2.2 Redshift

Redshift is the shift of the spectrum of an astronomical object towards longer wavelengths. It is defined as the ratio between the observed wavelength  $\lambda_{\text{obs}}$  and the emitted or rest-frame wavelength  $\lambda_{\text{RF}}$  of a distant object:

$$1 + z = \frac{\lambda_{\text{obs}}}{\lambda_{\text{RF}}}. \quad (1.15)$$

Generally, the change in wavelength can be due to three causes:

- **Cosmological expansion:** Under an expanding universe, the energy of photons traveling through it will decrease. The energy density of photons decreases as  $\rho \propto a^{-4}$ , meaning that photons lose energy as the scale factor  $a$  increases. Following the definition of redshift, we get:

$$1 + z_{\text{exp}} = \frac{1}{a(t_e)} \quad (1.16)$$

where  $a(t_e)$  is the scale parameter at the time the light was emitted.

- **Peculiar velocity:** This component arises from the peculiar velocity of the object due to the relative motion within its local environment. It can be interpreted as a relativistic Doppler effect:

$$1 + z_D = \sqrt{\frac{1 + v/c}{1 - v/c}}, \quad (1.17)$$

where  $v$  is the peculiar velocity in the radial direction.



- **Gravitational redshift:** GR shows that photons leaving a gravitational well will suffer a decrease in energy and therefore an increase in redshift  $z_{\text{grav}}$ . This effect is only relevant under the presence of heavy gravitational fields and is mostly negligible in cosmology.

All these three need to be taken into account to obtain the effective redshift perceived by the observer:

$$1 + z = (1 + z_D)(1 + z_{\text{grav}})(1 + z_{\text{exp}}), \quad (1.18)$$

but it is precisely the cosmological expansion redshift  $z_{\text{exp}}$  that helps us use redshift as a measurement of expansion, when neglecting the other contributions. It allows us to measure how much the universe has expanded since the emission of the radiation, provided the original wavelength is known.

Given that redshift can be directly measured, it is broadly used in modern cosmology. The only thing needed is a cosmological model that can translate the measurements of the scale parameter  $a$  (through Eq. (1.16)) into a distances.

### 1.2.3 The Hubble law

In 1929, Hubble (Hubble, 1929) found an approximately linear relationship between the physical distance of nearby galaxies and their recession velocity:

$$z = H_0 d, \quad (1.19)$$

where  $H_0$  is the Hubble constant. We can relate this equation to the scale parameter by expanding the latter in powers of  $(t - t_0)$ . To first order,  $a(t) = a(t_0) + \dot{a}(t - t_0)$ . Remembering that we follow the convention where  $c = 1$ , we have  $(t - t_0) = d$ . Therefore, using Eq. (1.19) along with Eq. (1.16):

$$H_0 = \frac{\dot{a}(t_0)}{a(t_0)}. \quad (1.20)$$

It is common to extend this definition to any point in the universe evolution, such that the Hubble parameter at any time can be defined as:

$$H(t) = \frac{\dot{a}(t)}{a(t)}. \quad (1.21)$$

### 1.3 The Friedman-Lamaître-Robertson-Walker metric

The Cosmological Principle states that the universe, when viewed on a sufficiently large scale, is uniformly isotropic and homogeneous. A general metric that fulfills this requirement while also allowing for a time-dependent spatial component (to integrate the observed expansion) is the Friedman-Lamaître-Robertson-Walker (FLRW) metric:

$$ds^2 = -dt^2 + a^2(t) \left[ \frac{dr^2}{1 - kr^2} + r^2 d\Omega^2 \right] \quad (1.22)$$

Written here in spherical spatial coordinates ( $r, d\Omega = d\theta^2 + \sin^2 \theta d\phi^2$ ). The scale factor  $a(t)$  represents the normalized size of the universe, accounting for its expansion or contraction. The curvature parameter  $k$  defines the curvature of the universe, being open for  $k < 0$ , closed for  $k > 0$  or flat for  $k = 0$ .

One can solve the Einstein equations (Eq. (1.8)) for the specific case of FLRW obtaining the Friedmann equations:

$$\left( \frac{\dot{a}}{a} \right)^2 = \frac{8\pi G}{3} \rho - \frac{k}{a^2} + \frac{\Lambda}{3} \quad (1.23)$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} (\rho + 3P) + \frac{\Lambda}{3} \quad (1.24)$$

The first equation relates the curvature and density with the expansion rate given by  $\dot{a}$ , we usually use the Hubble parameter  $H = \frac{\dot{a}}{a}$  to describe the expansion. The second equation describes the acceleration of this expansion, which is always negative for ordinary matter (with positive pressure and density). This behavior changes when we include the cosmological constant  $\Lambda$ , behaving as a perfect fluid with negative pressure.

Using the FLRW metric, along with the assumed stress-energy tensor in Eq. (1.9) and its conservation, we arrive at the conservation equation under the FLRW metric:

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + P) = 0 \quad (1.25)$$

The use of the Friedmann equations (Eq. (1.23) and Eq. (1.24)) and the conservation equation will describe the evolution of the homogeneous universe. In the definition of the stress-energy tensor in Eq. (1.9), we have only considered a single ingredient in the composition. In the next Section, we will see how different kinds of ingredients behave under the FLRW metric.

## 1.4 Cosmic ingredients

The evolution dictated by the Friedmann equations will depend on the content of the universe and the physical properties of this content. Our current understanding of the cosmos includes a variety of ingredients, each behaving distinctly and dominating the evolution at different cosmic times.

The single parameter that better characterizes a given fluid is the equation of state. Following thermodynamics, it is possible to model the dependence of pressure  $P$  and energy density  $\rho$  for a given fluid in the following way:

$$P = w\rho, \quad (1.26)$$

where  $w$  is the equation of state parameter. For a fluid composed of relativistic particles, we have  $w = \frac{1}{3}$ , while for ordinary non-relativistic matter, we have  $w = 0$ .

Under the assumption of constant  $w$ , it is possible to solve the conservation equation (Eq. (1.25)):

$$\rho = \rho_0 \left( \frac{a_0}{a} \right)^{3(1+w)} \quad (1.27)$$

where the subscript 0 defines a particular moment in time. This equation determines the evolution of density depending on the equation of state parameter. In an expanding universe, different fluids will dilute at different rates, leading to the dominance of some fluids at different times. Generally, we have three different behaviors:

- $w = 0$ : This is the behavior of pressureless non-relativistic matter, including both Cold Dark Matter (CDM) and baryons at large scales. CDM is a type of dark matter composed of non-relativistic particles that interact weakly with electromagnetic radiation and ordinary matter. It plays a crucial role in the formation of structure, since it can interact gravitationally with baryons, and explains the observed distribution of matter. With the expansion of the universe, it dilutes as  $\rho \propto a^{-3}$ , meaning it dilutes with the increase of volume.
- $w = \frac{1}{3}$ : This positive pressure behavior is carried out by relativistic particles, being the two most common representatives of it radiation in the form of photons and relativistic neutrinos at early times. With the expansion of the universe, its energy density decays as  $\rho \propto a^{-4}$ , meaning that it dilutes with the increase of volume, but also suffers redshifting. Most of the radiation observed today comes from free photons coming from the Cosmic Microwave Background (CMB), only contributing by a small factor to the total energy density.

- $w = -1$ : This behavior would correspond to the cosmological constant if it were to be included in this formalism. Its energy density is constant as the universe expands ( $\rho \propto a^0$ ), and therefore it will become dominant in later times. In the  $\Lambda$ CDM model, DE has this behavior, although more complex models include a DE component that behaves as  $w < -1/3$ , even having a equation of state parameter that evolves with time.

Our universe consists of a mixture of different components behaving as the three described above. Therefore, the total energy will consist of a mixture of these different components.

We can rearrange the Friedman equation (Eq. (1.23)) in the following way:

$$1 = \frac{8\pi G}{3H^2} \left( \rho - \frac{3k}{8\pi G a^2} + \frac{\Lambda}{3H^2} \right). \quad (1.28)$$

Using this expression, we can include in our formalism the curvature term as a density component with  $\rho_k = \frac{-3k}{a^2 8\pi G}$  and  $\rho_\Lambda = \frac{\Lambda}{3H^2}$  and then decompose the total density  $\rho$  as the sum of these components ( $\rho = \rho_k + \rho_r + \rho_m + \rho_\Lambda$ , where  $r$  and  $m$  account for radiation and matter). The preceding term in Eq. (1.28) is the critical density  $\rho_c = \frac{3H^2}{8\pi G}$ , with all this the Friedman equation can then be reformulated as follows:

$$1 = \frac{\rho_r}{\rho_c} + \frac{\rho_m}{\rho_c} + \frac{\rho_k}{\rho_c} + \frac{\rho_\Lambda}{\rho_c} = \Omega_r + \Omega_m + \Omega_k + \Omega_\Lambda, \quad (1.29)$$

where we used the density parameter  $\Omega_{(a)} \equiv \frac{\rho_{(a)}}{\rho_c}$ , usually defined at present time and whose evolution depends on the equation of state parameter (as defined in Eq. (1.27)). In this way, we can include the evolution as:

$$H^2 = H_0^2 \left( \Omega_r a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda \right). \quad (1.30)$$

Observations in the CMB, combined with Baryon Acoustic Oscillations (BAO) measurements lead to the conclusion that the curvature of the universe is consistent with 0 ( $\Omega_k = 0.001 \pm 0.002$  from Planck Collaboration et al. (2020)). Fig. 1.1 shows the evolution of energy density for the three most relevant species in the  $\Lambda$ CDM model. At early times, radiation was the dominant species, but since it decays faster than matter density with the expansion, around  $z \sim 3900$ , the energy density of matter becomes higher, entering a period of matter domination. Very recently, around  $z \sim 0.30$ , the energy density of matter decays so much that the energy density of DE (whose evolution is constant with the expansion) surpasses it, entering a period of DE domination.

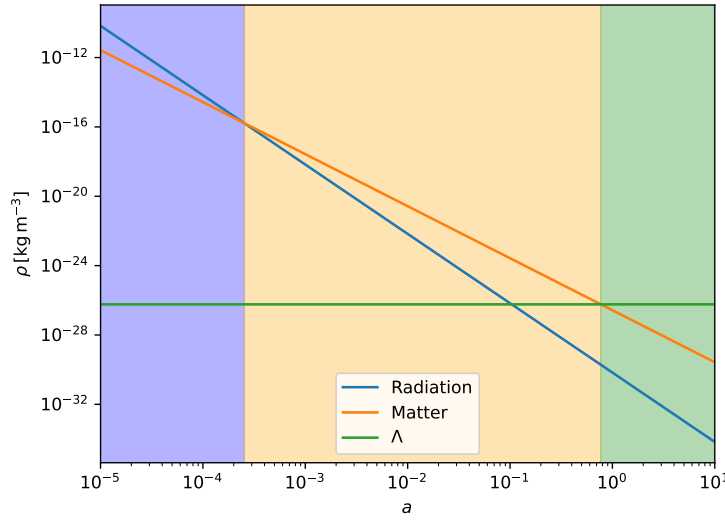


Figure 1.1: Energy content of the universe as a function of the scale factor  $a$ . The shaded regions mark the species that dominate at each stage of the expansion history. At early times, radiation was dominant; at  $z \sim 3900$ , the matter-radiation equality occurred; and from then on, matter was dominant. Very recently, at  $z \sim 0.30$ , DE began to dominate.

## 1.5 Thermal history

In this Section, we will briefly describe the evolution of the universe from the Big Bang to the present day to provide context for the following parts of the thesis.

After the Big Bang, the universe was an extremely dense plasma composed of quarks, gluons, leptons, photons and other fundamental particles. With expansion and cooling, the different particles of the Standard Model of particle physics began to combine, forming increasingly complex systems. This process was possible because the various particles that made up the primordial plasma started decoupling from each other.

To understand the concept of decoupling, one must consider two opposing elements: the interaction rate between particles and the expansion rate. When the interaction rate is higher than the expansion rate, particles interact faster than the universe expands and thus remain in thermal equilibrium. But if the universe expands rapidly enough, these particles cannot keep up and cease to interact.

With this mechanism, as the universe expands and temperature decreases, different particles decouple and become free from interaction. Shortly after the Big Bang, the universe was dominated by radiation and therefore its expansion evolved as  $a(t) \propto t^{1/2}$ . In this context, the first known particles to undergo

decoupling were neutrinos, when the temperature reached approximately 1 MeV<sup>1</sup>. This event generated a neutrino background similar to the photon CMB, although this is still to be measured.

When the temperature dropped to around 100 keV, the process of Big Bang Nucleosynthesis (BBN) occurs, where light elements are generated from the plasma of protons and neutrons. This is responsible for the formation of most of the universe's helium, along with small fractions of deuterium and lithium.

The moment when the temperature dropped to approximately 0.75 eV marks the matter-radiation equality, where the energy density of matter surpasses that of radiation and begins to dominate the evolution of the universe, and now its expansion evolved as  $a(t) \propto t^{2/3}$ . This event happened at redshift  $z \sim 3000$ .

At  $z = 1100$ , recombination occurs, where the temperature is low enough for electrons and protons to combine and form the first atoms. Due to the drastic reduction in the number of free electrons, photons eventually decoupled and traveled freely through the universe, forming what we now know as the CMB.

After this event, the slow gravitational collapse of baryonic matter along with the already collapsing dark matter begins. Dark matter had barely interacted with the rest of the matter in the past (its decoupling is expected to be much earlier), and its gravitational interaction had been very small compared to the other forces involved in the primordial plasma. This gravitational collapse led to the formation of the structure that we see in our universe.

This structure formation eventually led to the formation of galaxies, whose stars shone for the first time<sup>2</sup>. The light from these galaxies began to gradually ionize the hydrogen in the Intergalactic Medium (IGM), causing the universe to transition from being almost completely neutral to having a neutral fraction of around  $10^{-5}$  at  $z \sim 3$ .

Finally, much more recently, around  $z \sim 0.3$ , the presence of DE becomes dominant in the energy density of the universe, initiating a phase of accelerated expansion where  $a(t) \propto \exp(Ht)$ . This accelerated behavior was first confirmed using type-Ia supernovae at the end of the 20<sup>th</sup> century.

---

<sup>1</sup>We use units of energy here for convenience, the Boltzmann constant  $k_B$  provides a direct conversion between these two units:  $1 \text{ eV} = 1.16 \cdot 10^4 \text{ K}$

<sup>2</sup>The James Webb Space Telescope (JWST) is observing galaxies up to  $z = 13.2$ .



# ORIGIN AND EVOLUTION OF INHOMOGENEITIES

Our universe is clearly not completely homogeneous; we observe voids, clusters of galaxies, and a distinctive filamentary structure. The cosmological model described in the previous Chapter is not capable of explaining these patterns; in fact, it cannot describe any inhomogeneity at all.

In short, we need to add to our model a source of inhomogeneities, which is understood today to be quantum fluctuations at microscopic scales. Following the theory of inflation, these fluctuations grew exponentially during the early stages of the universe's evolution in the inflationary epoch. This allowed for the preservation of the isotropy and homogeneity of the observable universe, as we will see later.

In this Chapter, we will explore how the structure of the universe is formed. We will start by giving a short description of inflation, the source of inhomogeneities. Then, we will explain how linear theory allows us to study, to first order, how the inhomogeneities evolve with time. We will follow the derivations presented in Dodelson and Schmidt (2020).

## 2.1 Inflation

The current consensus is that small inhomogeneities in the early phases of the universe were caused by quantum fluctuations. However, this does not solve other problems in the Friedman-Lemaître-Robertson-Walker (FLRW) framework described in Chapter 1, mainly the flatness and the horizon problems:

- **Horizon problem:** According to the cosmological principle, the universe is homogeneous at large scales, and observations confirm this principle.



However, it implies that causally disconnected regions of spacetime have shared some sort of information. A period in the early universe of inflation, where the comoving horizon<sup>1</sup> shrinks, could explain this behavior by pushing causally connected regions outside the horizon.

- **Flatness problem:** Current measurements of the universe at all its stages show that its curvature is minimal, if not 0. This is puzzling because small variations in the curvature at early times should have grown drastically over time. Again, a period of inflation could explain this behavior because any existing curvature would have been smoothed out with the expansion.

The introduction of a period of rapid expansion in the very early universe can solve these problems. However, the precise mechanism behind this is still unclear. In general, this behavior occurs when the universe is dominated by a fluid with  $w < -1/3$ .

The inflaton  $\phi$  is proposed as the scalar field responsible for this inflationary period, with an energy density:

$$\rho = \frac{\dot{\phi}^2}{2} + V(\phi), \quad (2.1)$$

where  $V(\phi)$  is the potential. The equation of state for this fluid is:

$$w = \frac{\frac{\dot{\phi}^2}{2} - V(\phi)}{\frac{\dot{\phi}^2}{2} + V(\phi)}, \quad (2.2)$$

and therefore accelerated expansion can be achieved if the potential dominates over the kinetic energy, known as the slow-roll limit.

Inflation theory also provides a mechanism for the creation of initial perturbation. It magnifies quantum fluctuations, pushing them beyond the horizon and freezing them. Once the fluctuations re-enter the horizon they will evolve again. We will see how this mechanism works in Section 2.2.

Generally, additional conditions are imposed on the duration of the inflationary period to meet the observations of causally connected regions. For regions that are in causal contact today to have been in contact in the past, approximately 60 e-folds are required before the end of inflation.

---

<sup>1</sup>The horizon delineates the region of spacetime from which a given point can receive information, thereby defining the points that are causally connected to it.

## 2.2 Linear evolution

If we consider small perturbations to the FLRW metric described in Eq. (1.22), we need two fields that depend on space and time to describe the perturbed metric:

$$ds^2 = a^2(\eta) \left[ -(1 + 2\Psi)d\eta^2 + (1 - 2\Phi)\delta_{ij}dX^i dX^j \right], \quad (2.3)$$

where the fields  $\Phi$  and  $\Psi$  describe the perturbation,  $\delta_{ij}$  represents the Kronecker delta, and  $\eta = \int_0^t \frac{dt'}{a(t')}$  is the conformal time.  $\Psi$  is related to the Newtonian gravitational field, while  $\Phi$  is related to perturbations in curvature, with  $\Psi = \Phi$  in the absence of anisotropic stress. We will work under this assumption.

### 2.2.1 Evolution of the perturbations

Following the evolution of the potential  $\phi$  involves dealing with complex differential equations, but in Fourier space, each of the modes evolves independently, allowing the equations to be handled in a much simpler way.

The amplitude of a Fourier mode for a generic function  $f$  in configuration space is defined as:

$$\hat{f}(\mathbf{k}) = \int f(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}} d^3\mathbf{r}. \quad (2.4)$$

Similarly, we can use the inverse Fourier transform to recover the perturbations in real-space from the ones in Fourier-space:

$$f(\mathbf{r}) = \frac{1}{(2\pi)^3} \int \hat{f}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}} d^3\mathbf{k}. \quad (2.5)$$

It is common when using physical fields to use the same notation for functions in configuration space and Fourier space (for example, writing the density as both  $\delta(\mathbf{r})$  and  $\delta(\mathbf{k})$ ).

We will now use perturbations in Fourier space to describe how the evolution of the potential changes depending on the size of these modes and the properties of the species. Assuming a perfect fluid with adiabatic perturbations, the evolution of the potential  $\Phi$  is determined by:

$$\Phi'' + 3(1 + w)aH\Phi' + wk^2\Phi = 0, \quad (2.6)$$

for this equation we can consider two different behaviors:

- **Super-horizon modes:** Super-horizon modes are the ones where the third term in the equation can be neglected, due to  $k \ll aH$ , this means that fluctuations are on scales larger than the observable horizon. There are two

solutions for the remaining equation, one rapidly decaying with time and a constant solution, in which we will focus. This means that modes that are super-horizon remain "frozen" until they eventually get into the horizon with the expansion.

- **Sub-horizon modes:** When the size of the modes is comparable to the size of the horizon, we can no longer ignore the third term.

Focusing in the sub-horizon modes, we can consider the evolution of a radiation dominated universe, therefore having  $w = 1/3$ , for this case, the Eq. (2.6) becomes:

$$\Phi'' + \frac{4}{\eta}\Phi' + \frac{k^2}{3}\Phi = 0, \quad (2.7)$$

where we used  $a \propto \eta$ , and  $H \propto a^{-2}$  in a radiation-dominated universe. The dominant solution for this equation is of the type  $\Phi \propto \cos(k\eta/\sqrt{3})/(k\eta)^2$ , which means decaying oscillations.

The case of a matter-dominated universe is much simpler, since  $w = 0$ , we have:

$$\Phi'' + 3aH\Phi' = 0, \quad (2.8)$$

yielding a dominant solution with a constant value. For a matter-dominated era, both super-horizon and sub-horizon modes are frozen.

Finally, in the case of a Dark Energy (DE)-dominated universe, the growth of structure is slowed down for all relevant modes since they are all inside the horizon.

### 2.2.2 Evolution of species

Now we are going to delve into the evolution of the different species that make up the universe. Here we are going to work with the different density fluctuations, instead of the gravitational potential. We define the overdensity for species  $i$  as:

$$\delta_i = \frac{\delta\rho_i}{\bar{\rho}_i}. \quad (2.9)$$

It is then possible to describe the evolution of these density perturbations through the combination of the continuity and Euler equations in Fourier space with the Poisson equation, yielding:

$$\ddot{\delta}_i + (1 - 3w) \cdot H\dot{\delta}_i + \left( \frac{c_s^2 k^2}{a^2} - 4\pi G\bar{\rho} \right) \delta_i = 0, \quad (2.10)$$

where we use the sound speed in the fluid  $c_s$ . Note that this equation takes into account the Hubble drag (through  $H$ ), the pressure of the fluid in question (through  $c_s$ ), and the effects of gravity. There exists a critical scale  $k_J = \sqrt{4\pi G \bar{\rho} / c_s^2}$ , the Jeans scale, defining the scales beyond which perturbations cannot grow due to pressure effects. Depending on the species, this scale will be larger or smaller: for radiation, since  $c_s$  is very large, only the largest scales can grow.

If we want to follow the evolution of pressureless dark matter, we can ignore the speed of sounds. In this case the previous equation can be rearranged by using the Friedman equation and the substitutions  $y = a/a_{\text{eq}}$  (where  $a_{\text{eq}}$  is the scale factor at the matter-radiation equality), to get the Mészáros equation (Meszaros, 1974):

$$\frac{d^2 \delta_c}{dy^2} + \frac{2+3y}{2y(y+1)} \frac{d\delta_c}{dy} - \frac{3}{2y(y+1)} \delta_c = 0. \quad (2.11)$$

This equation express how matter sub-horizon modes evolve with time, and from it we can obtain different solutions for different epochs. In radiation domination, we have logarithmic growth ( $\delta_c \propto \ln a$ ); in the case of matter domination, we have  $\delta_c \propto a$ ; and in the case of DE domination, we have  $\delta_c \propto \text{const}$ .

The matter growth of structure is usually parameterized with the growth factor  $D(z)$  as follows:

$$\delta(\mathbf{k}, t) = \frac{D(z)}{D(z_0)} \delta(\mathbf{k}, t_0), \quad (2.12)$$

where its particular value will depend on the cosmological model used. This expression can only be used during matter domination or DE domination epochs, in the case of radiation domination eras, there would be a dependence depending on whether the specific modes are inside or outside the horizon.

### 2.2.3 Photon-baryon oscillations

In the early universe, photons and baryons are coupled due to Compton scattering and can effectively be treated as a single fluid.

For this case, density perturbations follow the equation of an oscillator with frequency (Eisenstein et al., 2007):

$$\nu = c_s k = \frac{k}{\sqrt{3 \left(1 + \frac{3\rho_b}{4\rho_\gamma}\right)}}. \quad (2.13)$$

The acoustic waves produced by these oscillations leave an imprint in the distribution of baryons and photons, known as Baryon Acoustic Oscillations (BAO).

The understanding of this distribution is as follows: in a region with overdensity compared to its surroundings, a spherical pressure wave is released outward due to the extra pressure from the overdensity. This pressure wave would travel indefinitely if it weren't for the moment in the expansion when photons and baryons decouple. In this new situation the pressure drops and the waves freeze. While the photons stream away, the baryons become fixed in a spherical shell given by (Hu and White, 1996):

$$r_d = \int_0^\eta d\eta' c_s = \int_{z_d}^\infty \frac{c_s(z)}{H(z)} dz, \quad (2.14)$$

where  $z_d \simeq 1020$  is the redshift when the decoupling occurred. This value is found to be around  $r_d \simeq 150$  Mpc (Planck Collaboration et al., 2020).

After decoupling, baryons will evolve similarly to dark matter (that stayed at the centers of the overdensities). Gravitational attraction will make baryons fall into the original overdensities, leaving an imprint in the overall matter density with a characteristic scale set by  $r_d$ .

### 2.2.4 The two point correlation function

One of the most relevant statistics in current cosmology that allow us to understand the distribution of the different clustering modes is the two point correlation function and the associated power spectra. The correlation function measures how a given field is clustered as a function of distance. For a density field  $\delta$  the correlation function is defined as:

$$\xi(\mathbf{x}, \mathbf{y}) \equiv \langle \delta(\mathbf{x})\delta(\mathbf{y}) \rangle = \xi(r = |\mathbf{x} - \mathbf{y}|), \quad (2.15)$$

where in the last equality we used the cosmological principle, this function cannot depend on the orientation of the two points. The Fourier transform of the two-point function is called the power spectrum  $P(k)$ , and its use is even more widespread in cosmology:

$$\xi(r) = \frac{1}{2\pi^2} \int dk k^2 P(k) \frac{\sin(kr)}{kr}. \quad (2.16)$$

The power spectra measures the strength of density fluctuations in the universe as a function of the scale of these fluctuations, helping us to understand how the given tracer of the underlying density field is distributed in space.

The theory of inflation predicts a primordial power spectrum of the form:

$$P(k) = A_s \left( \frac{k}{k_0} \right)^{n_s}, \quad (2.17)$$

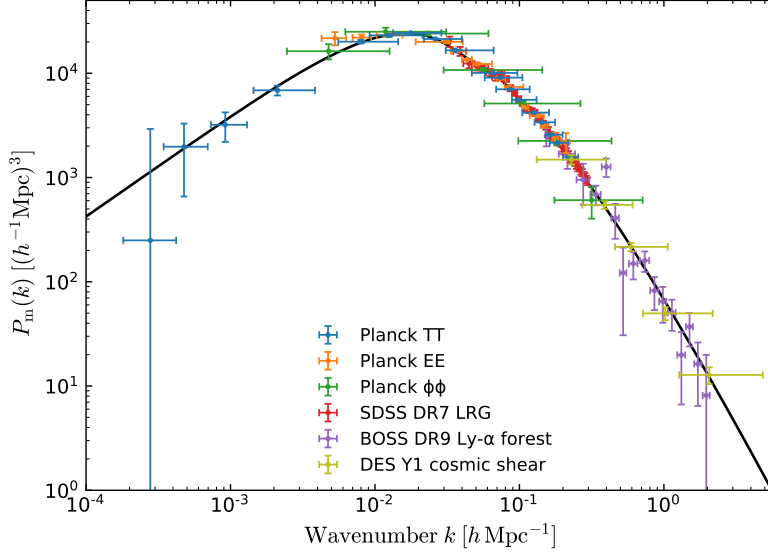


Figure 2.1: The (linear theory) matter power spectrum fitted from multiple cosmological probes. The data markers correspond to inferences at  $z = 0$  based on different clustering measurements.

where  $A_s$  is the amplitude,  $k_0 = 0.05 \text{ Mpc}^{-1}$  the pivot point, and  $n_s$  the spectral index, recently measured by Planck as  $n_s = 0.965 \pm 0.004$  (Planck Collaboration et al., 2020).

The evolution of the primordial density field towards the linear power spectrum of today is parametrized by the transfer function  $T(k, z)$  and the growth  $D^2(z)$  as:

$$P(k, z) = T^2(k) \cdot \frac{D^2(z)}{D^2(z=0)} \cdot P_0(k), \quad (2.18)$$

where  $P_0$  the primordial power spectrum.  $P_0$  takes into account the impact of the interplay between the modes and the horizon (as shown in Section 2.2.2), and the growth of structure taking place at later times is governed by the parameter  $D$ . The amplitude of the power spectrum is defined from observations, usually considering density fluctuations in a sphere of radius  $8 \text{ Mpc}/h$ , a quantity defined as  $\sigma_8$ , being the measurement from the Planck collaboration  $\sigma_8 = 0.811 \pm 0.006$ . In Fig. 2.1 we show the measured power spectrum from clustering and Cosmic Microwave Background (CMB) probes, alongside with the linear theory fit, corresponding to the one in Eq. (2.18).

#### 2.2.4.1 Angular two-point correlation functions

In photometric surveys, it is common to use the two-point correlation function in its angular form, Usually because the redshift position of the galaxies is not possible to be accurately determined. Generally, techniques are used to determine

the approximated redshift of the galaxies, and then they are grouped into more or less thicker redshift bins. In these bins, the 2D angular correlation function is then measured, only requiring the angular position of the objects. Due to isotropy, this function will only depend on the absolute value of the angular separation between galaxies,  $\theta$ .

The Fourier transform of the angular two-point correlation function is the angular power spectrum. For two fields A and B, defined as projections of the 3D fields  $a$  and  $b$ , the angular power spectrum has the form:

$$C_{AB}(\ell) = \int d\chi \frac{q_A(\chi)q_B(\chi)}{\chi^2} P_{ab} \left( \frac{\ell + 1/2}{\chi}, z(\chi) \right), \quad (2.19)$$

where  $P_{ab}$  is the three-dimensional power spectrum relative to the two fields  $a$  and  $b$ , computed for the wave number  $k = \ell/\chi$ , where  $\ell$  indicates the multipole, and  $q_\chi$  are the window functions. The window function act as weights that quantifies the number of galaxies expected to be observed if the universe was unclustered, capturing both geometrical and observational effects (Karim et al., 2023).

### 2.2.5 Non-linear evolution

All the theory deployed above has a validity range limited by the assumption of linearity. Once we enter regimes where  $\delta > 1$ , this approximation becomes insufficient to describe all the physics involved. Large fluctuations lead to the formation of virialized dark matter structures called halos. These halos will interact with other nearby halos and form much more complex structures.

Due to advances in computing power, we can simulate many of these effects at various levels. In N-body simulations, the gravitational interaction between objects is simulated, helping us to understand the formation of structure at all scales with more precision. However, baryonic effects are not included and therefore the clustering will not be accurate at very small scales. These effects are simulated in hydrodynamical simulations, and these simulations allow us to study the evolution of baryonic gas more accurately, although they are computationally expensive and can only cover small volumes.

Other methodologies, as Lagrangian Perturbation Theory (LPT), allow for the generation of non-linear physical matter overdensities while keeping the computational needs at minimum. LPT is able to capture non-linear aspects of the matter overdensity by carrying out low-order perturbation theory calculations in the displacement field. In LPT, the linear overdensity is used to predict the Lagrangian displacement of a set of massive test particles, then the non-linear density field is given by the density of the displaced test particles. This method

has been used in the past to generate mock galaxy catalogs (Manera et al., 2013; Chuang et al., 2015; Manera et al., 2015). In Chapter 5 we will revisit LPT in the context of cosmological simulations.





## COSMOLOGICAL PROBES

In recent decades, we have entered the era of precision cosmology, handling a wealth of data that increasingly resembles that of particle physics. This is due to the improvement of the technology in cameras and telescopes, allowing us to obtain more precise and abundant measurements than ever before.

These measurements exist for a wide range of tracers at different moments in the evolution of the universe. In this Chapter we will detail some of the most relevant for the core Chapters of this thesis.

We will begin by discussing the Cosmic Microwave Background (CMB), the oldest light we can observe in the night sky and how it brings us closer to the Big Bang. We will then move on to measurements of clustering based on galaxy positions, followed by measurements of neutral hydrogen absorption in the Lyman-alpha forest, and finally conclude with a brief introduction to weak lensing.

### 3.1 The Cosmic Microwave Background

The CMB is the cosmological probe that allows us to travel furthest back in time with our measurements. It is composed of free photons released after the time of recombination, at  $z_r \simeq 1090$ . As mentioned before in Section 2.2.3, before this moment, photons and baryons were coupled. As the universe cooled and reached a temperature of  $T \simeq 3300K$  (equivalent to  $\sim 0.3eV$ ) around 380 000 years after the Big Bang, the number of neutral atoms surpassed the number of ionized atoms, in an epoch known as recombination. After this event, the number of free electrons decreased, increasing the mean free path of photons, effectively making the universe transparent.

From that moment on, radiation cooled with the expansion of the universe until reaching the temperature of  $T_0 = 2.72548 \pm 0.00057K$  measured currently (Fixsen, 2009). These observed photons make up a 2D surface called the "surface

of last scattering," and multiple maps of this radiation have been explored since its first observation in 1964 (Dicke et al., 1965; Penzias and Wilson, 1965). The Cosmic Background Explorer (COBE) confirmed that the CMB has a blackbody spectrum (Mather et al., 1990), and detected for the first time anisotropies in it (Smoot et al., 1992). However, COBE did not have sufficient resolution to detect the acoustic oscillations described in Section 2.2.3.

The next large CMB survey was the Wilkinson Microwave Anisotropy Probe (WMAP; Bennett et al., 2003), which was able to resolve the acoustic peaks (allowing for cosmological constraints) and measure the anisotropies with higher resolution. More recently Planck (Planck Collaboration et al., 2020) provided the most accurate measurement of the CMB temperature and polarization anisotropies to date.

The small fluctuations observed by these surveys are at the order of one part in  $10^5$ , and are understood as remnants of the quantum fluctuations present at the time of recombination, which eventually lead to the formation of halos and galaxies.

Typically, CMB surveys focus on the angular power spectrum. This allows us to observe the oscillations of the baryon-photon fluid described in Section 2.2.3, and provide constraints for several cosmological parameters.

There are also secondary anisotropies that are measured in the CMB, caused by effects on the CMB light while it travels through the universe. For example, the effect of time-varying gravitational potentials in the late universe will affect the wavelength of photons (Sachs-Wolfe effect; Sachs and Wolfe, 1967); or high-energy electrons in galaxy clusters will affect the wavelength of photons via inverse Compton scattering (Sunyaev-Zel'dovich; Sunyaev and Zeldovich, 1970; Sunyaev and Zeldovich, 1980).

## 3.2 Galaxy clustering

Perhaps the most obvious way to study our universe is to locate and map the galaxies within it. This simple task has been perfected over the years, allowing the observation of millions and millions of objects in the sky, and still today it is one of the most widely used methods to study the evolution and components of our universe.

There are two types of surveys that allow us to generate maps of galaxies, photometric and spectroscopic. Photometric surveys involve generating high-quality images of the night sky, which allows for the subsequent precise identification of

the position of galaxies in them. This method enables the acquisition of a generally very high number of objects with ease, but suffers from the drawback of not being able to accurately identify the redshift to each galaxy. To estimate this, different filters are used to "photograph" the sky at different wavelengths. By looking at the flux in the different bands, it is possible to estimate their redshift with some precision.

The other type of survey is spectroscopic surveys. In this case, each galaxy spectrum is extracted with higher accuracy, with the resolution usually depending on the science target, since generally the higher the resolution needed the fewer objects will be observed. As a result of this measurements, we obtain detailed information of each galaxy, specifically emission or absorption lines, which allow us to accurately identify the type of galaxy and its location in redshift. Of course, there are secondary effects complicating this task, such as the peculiar velocities of galaxies in clusters.

This method of measurement is more precise than the photometric method, but for logistical reasons, it is also much slower. Therefore, both types of surveys are still conducted today. For comparison, in the same generation of cosmological surveys, Dark Energy Spectroscopic Instrument (DESI) will observe around 40 million objects, while photometric surveys of the same generation, such as Euclid (Laureijs et al., 2011) and the Vera C. Rubin Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al., 2009), will map approximately a billion galaxies.

Galaxy clustering follows the clustering of the underlying dark matter, but not directly. There is a relationship between them, but it is a biased relationship, as galaxies will only form in the density peaks of dark matter. This relationship is modeled in linear order using the linear bias parameter  $b$  as follows (Kaiser, 1984):

$$\delta_g(\mathbf{x}) = b \delta(\mathbf{x}). \quad (3.1)$$

This parameter is generally considered to be dependent on the redshift and galaxy type, and will hold for large linear scales. The relation will break when entering non-linear scales where the matter fluctuations are not sufficiently small ( $\delta \gtrsim 1$ ).

### 3.2.1 Clustering in redshift space

When mapping objects in three dimensions in a survey, we usually use the redshift of the object to measure the radial distance. However, this measurement also includes effects from peculiar velocities, and hence include a systematic error in the measurement of the distance. If we only consider the Hubble flow, the

relation between the cosmological redshift  $z_{\text{cos}}$  and the comoving distance is:

$$r(z_{\text{cos}}) = \int_0^{z_{\text{cos}}} \frac{c \, dz}{H(z)}. \quad (3.2)$$

But, in the observed redshift of the object we also must take into account the redshifting due to the peculiar velocity of the galaxy. In this case, the observed redshift will have the form:

$$1 + z_{\text{obs}} = (1 + z_{\text{cos}}) \left( 1 - \frac{v_{\parallel}(\mathbf{r})}{c} \right)^{-1}, \quad (3.3)$$

where  $v_{\parallel}$  denotes the velocity of the object along the line of sight. We can then consider the distance in redshift space as (Zaroubi and Hoffman, 1993):

$$\mathbf{s} = \mathbf{r} + \frac{(1 + z_{\text{cos}})v_{\parallel}(\mathbf{r})}{H(z_{\text{cos}})} \hat{r}, \quad (3.4)$$

which would correspond to the estimated position of the galaxy directly from its redshift, without taking into account peculiar velocities. The difference between  $\mathbf{s}$  and  $\mathbf{r}$  is known as Redshift Space Distortions (RSD).

In the linear regime, we can express the density of galaxies in Fourier space as (Kaiser, 1987):

$$\delta_g^s(\mathbf{k}) = \delta_g(\mathbf{k}) + f\mu^2\delta(\mathbf{k}), \quad (3.5)$$

where  $\delta_g$  the density of galaxies and  $\delta$  the density of dark matter.  $f$  is the logarithmic growth rate, function of the growth rate (Peebles, 1980):

$$f \equiv \frac{d \ln D}{d \ln a}, \quad (3.6)$$

and  $\mu$  is the cosine of the line of sight angle.

Combining Eq. (3.1) and Eq. (3.5), we get the expression of the linear power spectrum of galaxies in redshift space:

$$P_g^s(k, \mu) = b^2(1 + \beta\mu^2)^2 P_m(k), \quad (3.7)$$

where  $\beta \equiv f/b$  and  $P_m$  is the linear power spectrum of the underlying dark matter field. Note that the galaxy power spectrum ceases to be isotropic because it is enhanced in the line of sight direction.

Although Redshift Space Distortions (RSD) can be thought of as a complication to the galaxy clustering, there is a lot of information we can learn from them and a lot of studies target it. It is common to express RSD measurements as the amplitude of the peculiar velocity field  $f\sigma_8(z)$ .

It is also common to measure the correlation function with RSD corrections using polynomial decomposition. To do this, we expand the power spectrum using Legendre polynomials:

$$P^s(k, \mu) = \sum_{\ell} P_{\ell}(k) \mathcal{L}_{\ell}(\mu) \quad (3.8)$$

$$P_{\ell}(k) = \frac{2\ell + 1}{2} \int_{-1}^1 d\mu P^s(k, \mu) \mathcal{L}_{\ell}(\mu). \quad (3.9)$$

Given that the Kaiser formula (Eq. (3.5)) only has terms up to order 4 in  $\mu$ , only the modes  $\ell = 0$  (monopole),  $\ell = 2$  (quadrupole), and  $\ell = 4$  (hexadecapole) will be non-zero:

$$P_{g,\ell=0}^s(k) = \left(1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2\right) b^2 P_m(k) \quad (3.10)$$

$$P_{g,\ell=2}^s(k) = \left(\frac{4}{3}\beta + \frac{4}{7}\beta^2\right) b^2 P_m(k) \quad (3.11)$$

$$P_{g,\ell=4}^s(k) = \frac{8}{35}\beta^2 b^2 P_m(k). \quad (3.12)$$

It is also possible to express the two-point correlation function in polynomials from the power spectrum:

$$\xi_{\ell}^s(s) = i^{\ell} \int \frac{k^2 dk}{2\pi^2} P_{\ell}^s(k) j_{\ell}(ks), \quad (3.13)$$

where  $j_{\ell}(x)$  is the spherical Bessel function of order  $\ell$ .

Finally, we can measure the correlation function (or the power spectrum) in wedges, defined as a simple average over a range of  $\mu$ :

$$\xi_{\mu_1}^{s,\mu_2}(k) \equiv \frac{1}{\mu_2 - \mu_1} \int_{\mu_1}^{\mu_2} d\mu \xi^s(k, \mu). \quad (3.14)$$

Throughout this Section, we have assumed the linear regime. However, this model becomes very limited when we try to model the non-linear scales. It is more common to try to model the behavior in these regions using perturbation theory (Desjacques et al., 2018). Similarly as described in Section 2.2.5, Lagrangian Perturbation Theory (LPT) models can also be used to most accurately describe the behavior of galaxies in the nonlinear regime (Maus et al., 2024).

### 3.2.2 BAO measurements from galaxy clustering

As we have seen before, the Baryon Acoustic Oscillations (BAO) signature imprinted by the oscillations of the photon-baryon fluid remains present in the matter distribution and therefore also in the distribution of galaxies. This scale  $r_d$  is known to be frozen since the decoupling and therefore can be used as a standard ruler to study the expansion of the universe at different epochs. It is even possible to obtain its physical value using parameters derived solely from CMB anisotropies, or from Big Bang Nucleosynthesis (BBN) abundance measurements alongside other cosmological parameters (Addison et al., 2013, 2018).

For objects at a similar redshift, we can measure the position of the BAO peak in the angular correlation function:

$$\theta_{\text{BAO}} = \frac{r_d}{(1+z)D_A(z)}, \quad (3.15)$$

allowing us to infer the angular diameter distance to these galaxies.

Similarly for the parallel direction, we can measure an excess correlation along the line of sight  $\Delta z_{\text{BAO}}$ , from which we can infer the expansion rate at the time,  $H(z)$ , through:

$$r_d = \int_z^{z+\Delta z_{\text{BAO}}} \frac{c}{H(z')} dz' \simeq \frac{c\Delta z_{\text{BAO}}}{H(z)}, \quad (3.16)$$

where we have reintroduced the factor  $c$ .

BAO from galaxy clustering has been measured multiple times since the first clear detection using Sloan Digital Sky Survey (SDSS) spectroscopic data (Eisenstein et al., 2005). The most recent by using the DESI Data Release 1 (DR1) spectroscopic sample of galaxy and quasars (DESI Collaboration et al., 2024a) and the Dark Energy Survey (DES) Y6 photometric galaxy sample (DES Collaboration et al., 2024). As we will see in the next Section, BAO can also be measured using another tracers, like the Lyman- $\alpha$  forest.

## 3.3 The Lyman- $\alpha$ forest

The Lyman- $\alpha$  forest is a pattern of absorption features caused by neutral hydrogen in the Intergalactic Medium (IGM), typically observed in the spectra of distant quasars ( $z > 2$ ), first detected in 1960 (Bahcall and Salpeter, 1965; Gunn and Peterson, 1965; Scheuer, 1965; Schmidt, 1965).

These features can be easily understood by considering the absorption of the Lyman- $\alpha$  transition line  $\lambda_\alpha = 1215.67 \text{ \AA}$  and the spectrum of a quasar that has been redshifted due to the expansion of space during its journey.

As the photons from the spectrum of a quasar travel through space, they become redshifted, meaning their wavelength increases over time. For wavelengths such that  $\lambda_{\text{rest}} \leq \lambda_\alpha$ , at some point during their journey, their wavelength will be exactly equal to  $\lambda_\alpha$ , and at that moment, they can be absorbed by neutral hydrogen if it is present.

Depending on the amount of gas, the absorption will be more or less pronounced, so that the observer on Earth will observe a reduction in the quasar flux at a wavelength  $\lambda_{\text{obs}} = (1 + z)\lambda_\alpha$ , where  $z$  is the redshift of the gas cloud. The effect of different gas clouds at different redshifts between the quasar and our galaxy generates an absorption pattern in the quasar spectrum, which is what we recognize as the Lyman- $\alpha$  forest.

The usefulness of the Lyman- $\alpha$  forest is that we can trace the dark matter field by looking at the spectra of these quasars. However, the relationship between the dark matter field and the received flux is not entirely trivial. The Lyman- $\alpha$  forest absorption will directly depend on the optical depth through  $F = \exp(-\tau)$ , where:

$$\tau = \int ds n_{\text{H I}}(z) \sigma(\nu), \quad (3.17)$$

where  $n_{\text{H I}}$  the neutral hydrogen number density and  $\sigma(\nu)$  the Lyman- $\alpha$  forest cross section at frequency  $\nu$ . This expression leads to the Gunn-Peterson optical depth (Gunn and Peterson, 1965):

$$\tau_{\text{GP}} = \frac{\pi e^2}{m_e c} f_\alpha \lambda_\alpha \frac{n_{\text{H I}}}{H(z)}, \quad (3.18)$$

where  $f_\alpha$  is the Lyman- $\alpha$  forest oscillator strength. We see how this relation will relate the flux of the quasar to the density of neutral hydrogen, however the peculiar velocities of the hydrogen gas is not taken into account. For a more complete expression of the optical depth see McQuinn (2016).

Since the fluctuations of the flux fraction  $F$  are not directly measurable, they must be obtained through the observed flux in the telescope,  $f$ :

$$\delta_F(\lambda) = \frac{f(\lambda)}{\bar{F}(\lambda)C(\lambda)} - 1, \quad (3.19)$$

where  $\bar{F}(\lambda)$  is the mean transmitted flux, and  $C(\lambda)$  the quasar continuum. This relationship is quite complicated because in cases where the signal-to-noise ratio is not particularly high (as in surveys where the observation of many targets is prioritized over overexposure to each target) it is difficult to generate an estimate for  $C(\lambda)$ .



We can derive the equations that describe the clustering of the Lyman- $\alpha$  forest in a similar way to what we have done in Section 3.2. At the linear level, the power spectrum of the Lyman-alpha transmitted flux fraction is given by (McDonald, 2003):

$$P_F^s(k, \mu) = (b_F + b_{\eta,F} f \mu^2)^2 P(k), \quad (3.20)$$

where  $b_{\eta,F}$  appears as we work with transmitted flux, which has a non-linear mapping to the distorted field of optical depth (Seljak, 2012; Arinyo-i-Prats et al., 2015).

Usually, both the auto-correlation of Lyman- $\alpha$  fluctuations and the cross-correlation of these fluctuations with quasars is computed. The auto-correlation measurement is performed using a weighted covariance estimator (Slosar et al., 2011):

$$\xi_A = \frac{\sum_{i,j \in A} w_i w_j \delta_i \delta_j}{\sum_{i \in A} w_i w_j}, \quad (3.21)$$

where  $A$  is the bin in  $(r_{\parallel}, r_{\perp})$  where the correlation is computed and  $w_i$  the weights that take into account the noise in each pixel (see Section 6.4.1). For the cross-correlation with quasars we use (Font-Ribera et al., 2012b):

$$\xi_A = \frac{\sum_{i,j \in A} w_i w_j \delta_i}{\sum_{i \in A} w_i w_j}. \quad (3.22)$$

Using data from the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al., 2013), the three-dimensional correlation function of absorption in the Lyman- $\alpha$  forest was measured for the first time in Slosar et al. (2011). Shortly after that, the first measurement of the BAO peak in the Lyman- $\alpha$  forest was presented (Busca et al., 2013; Kirkby et al., 2013; Slosar et al., 2013), using data from BOSS DR9 (Lee et al., 2013). These were followed by other BAO analyses using increasingly larger Lyman- $\alpha$  forest datasets from BOSS (Delubac et al., 2015; Bautista et al., 2017) and from the Extended Baryon Oscillation Spectroscopic Survey (eBOSS Dawson et al., 2016; de Sainte Agathe et al., 2019).

The precision of these BAO measurements was significantly improved with the measurement of the cross-correlation of quasars and the Lyman- $\alpha$  forest (Font-Ribera et al., 2014; du Mas des Bourboux et al., 2017; Blomqvist et al., 2019), and the final Lyman- $\alpha$  BAO measurement combining BOSS and eBOSS was presented in du Mas des Bourboux et al. (2020).

The last result from DESI Lyman- $\alpha$  is the measurement of BAO from the 3D correlations using DR1 data (DESI Collaboration et al., 2024b). In the top panel of Fig. 3.1, we show the Lyman- $\alpha$  auto-correlation from this dataset. The different colors show different orientations with respect to the line of sight, and the BAO

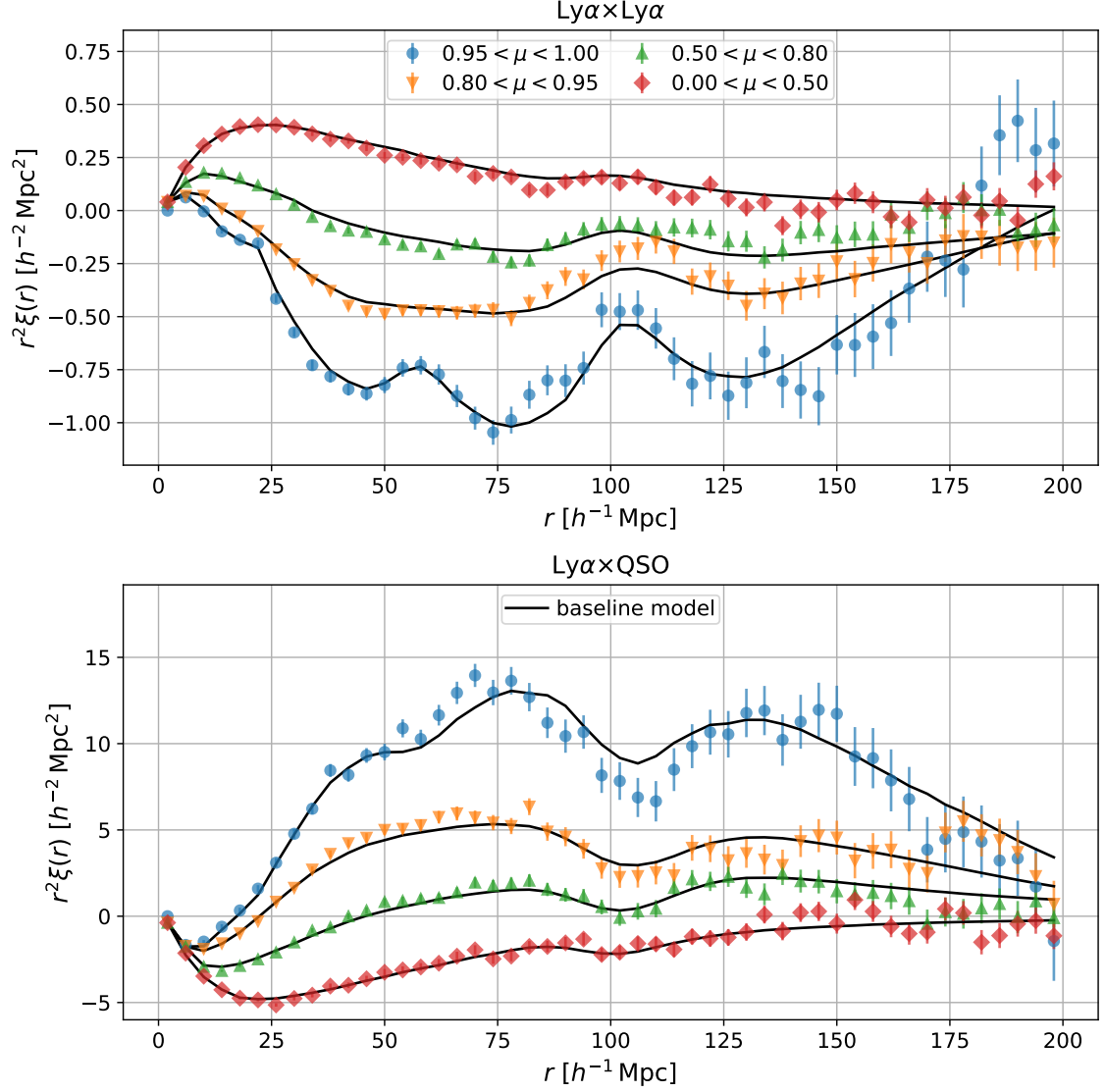


Figure 3.1: Top: Lyman- $\alpha$  auto-correlation along with the best fit model from the DESI DR1 BAO analysis. The different colors correspond to different orientations with respect to the line-of-sight, with blue correlations being close to the line-of-sight  $0.95 < \mu < 1$ . Bottom: Measured Lyman- $\alpha$  cross-correlation with quasars for the same survey. (DESI Collaboration et al., 2024b)

feature is clear in all of them. Absorption by heavier elements causes bumps at lower separations with high dependency on  $\mu$  that need to be correctly modeled in the analysis. In the bottom panel of the same Fig. 3.1, we show the Lyman- $\alpha$  cross-correlation with quasars from the same analysis. The correlation function is inverted with respect to the y-axis given that the Lyman- $\alpha$  flux has negative bias, and the BAO peak appears as a trough.

In Chapter 6, we will present the first Lyman- $\alpha$  catalog using DESI Early Data Release (EDR) data (Ramírez-Pérez et al., 2024), aimed to be used for 3D correlation studies, where the flux of different quasars is correlated with the aim of constraining the BAO scale (see Gordon et al. (2023) for details on the measurement of Lyman- $\alpha$  3D correlations using DESI EDR data). There are other alternative studies where only the 1D correlation is used, correlating fluxes of the same quasar spectra at different wavelengths, better constraining the small scales (Ravoux et al., 2023; Karaçaylı et al., 2024).

## 3.4 Weak Gravitational Lensing

Following the theory of General Relativity (GR), we know that particles propagate following the shortest path in spacetime. For massless particles like photons, this is defined as  $ds^2 = 0$ . Under the presence of mass the spacetime becomes distorted, and the path ceases to be a straight line. As a consequence of this, the shape and position of distant objects can be affected by the presence of matter between the observer and these objects. Sometimes, multiple images of the same galaxy or Einstein rings can even be generated.

In weak lensing, it is not these large effects that are of interest, but rather small variations in the appearance of millions of galaxies, which help us test large scale structure (LSS). These small variations come in the form of subtle differences in the shapes of galaxies measured photometrically. The two most typical ways to evaluate these effects are cosmic shear and galaxy-galaxy lensing:

- *Cosmic shear*: This effect occurs when the shape of distant galaxies is affected by the presence of a gravitational field between them and the observer, modifying their apparent shape and intensity. Studying the correlation between these galaxy shapes allows us to generate maps of the distribution of matter.
- *Galaxy-galaxy lensing*: In this case, the maps generated are correlated with galaxies at lower redshift, which form part of the halos sourcing the lensing effects.

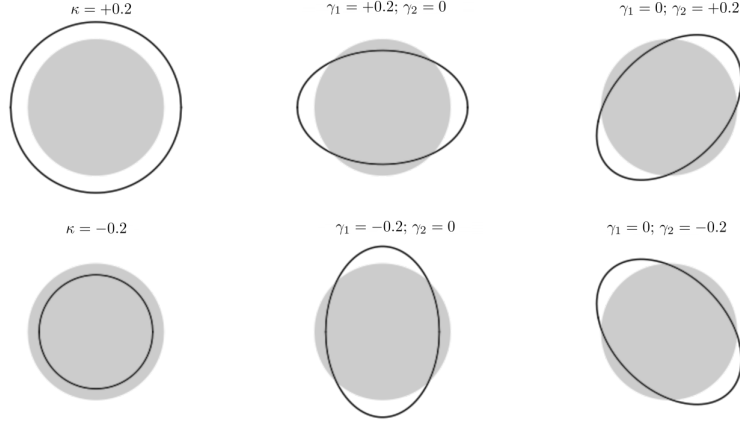


Figure 3.2: Effects of shear ( $\kappa$ ) and convergence (with components  $\gamma_{1,2}$ ) in the shape of an originally circular galaxy. The convergence magnifies the image, while the two shear components stretch the image in different directions (Bovy, 2024).

The effects of gravitational lensing on different cosmological observables are encoded in the so-called “lensing potential”, defined as (Bartelmann and Schneider, 2001; Lewis and Challinor, 2006)

$$\psi(\hat{\mathbf{n}}) \equiv -2 \int_0^{\chi_*} d\chi \frac{\chi_* - \chi}{\chi_* \chi} \phi_N(z(\chi), \chi \hat{\mathbf{n}}), \quad (3.23)$$

where  $\chi_*$  the comoving distance to the source.

The trajectories of photons are deflected by an angle  $\alpha \equiv \nabla_{\hat{\mathbf{n}}} \psi$ , and the shapes of background objects are distorted via the lensing distortion tensor  $\Gamma \equiv -\mathbf{H}_{\hat{\mathbf{n}}} \psi$ , where  $\nabla_{\hat{\mathbf{n}}}$  and  $\mathbf{H}_{\hat{\mathbf{n}}}$  are the gradient and Hessian operators on the sphere. The distortion tensor  $\Gamma$  is commonly decomposed into its spin-0 and spin-2 components, the convergence  $\kappa$  and shear ( $\gamma_1, \gamma_2$ ):

$$\Gamma \equiv \begin{pmatrix} \kappa + \gamma_1 & \gamma_2 \\ -\gamma_2 & \kappa - \gamma_1 \end{pmatrix}. \quad (3.24)$$

In Fig. 3.2, we show the effects of the fields  $\kappa$  and  $\gamma$  in the shape of an originally circular galaxy. The effect of the two effects combined transform a circular image of with radius  $r$  into an ellipse with semi-major and semi-minor axes (Bovy, 2024):

$$a = \frac{r}{1 - \kappa - \gamma}, \quad b = \frac{r}{1 - \kappa + \gamma} \quad (3.25)$$

Weak lensing probes directly the gravitational potential, and this is a powerful tool, lensing measurements combined with galaxy clustering allow to break degeneracies between cosmological parameters and biases, helping to constrain dark energy. In the last years a number of photometric surveys have been conducted, being the most recent results from the DES Y6 weak lensing sample.



# THE DARK ENERGY SPECTROSCOPIC INSTRUMENT

Dark Energy Spectroscopic Instrument (DESI) is the largest ongoing spectroscopic survey, continuing the efforts started in the BOSS and eBOSS collaborations. It operates on the Mayall 4-meter telescope at Kitt Peak National Observatory (DESI Collaboration et al., 2022) and is being led by the Department of Energy and in particular by the Lawrence Berkeley National Laboratory (LBNL).

The primary goal of this collaboration is to constrain Dark Energy (DE) models using Baryon Acoustic Oscillations (BAO), and to achieve this, it will measure more than 40 million objects over 14 000 square degrees, from nearby bright galaxies to distant quasars.

The DESI telescope consists of 5 000 fibers placed in the focal plane and distributed across 10 petals. Each fiber is controlled by a robotic positioner, allowing for quick and automatic positioning. The Guiding, Focusing & Alignment (GFA) cameras, the main contribution from IFAE and other Spanish institutions to the instrument, ensure that the robotic positioners can collect the light from the galaxies under optimum conditions.

The large number of fibers and the fast cadence provided by the automatic positioners allows DESI to measure up to 5 000 spectra every 20 minutes over a  $\sim 3^\circ$  field (DESI Collaboration et al., 2016b; Miller et al., 2023; Silber et al., 2023). Fibers carry light from the telescope into a separate room, where it is dispersed by ten spectrographs. Each of them with three cameras each (B, R, Z), covering different wavelength ranges.

Each of the three arms of the spectrograph covers a different wavelength region with some overlap: the B arm covers  $[3600, 5800]$  Å, the R arm covers  $[5760, 7620]$  Å and the Z arm covers  $[7520, 9824]$  Å. The Lyman- $\alpha$  forest is predominantly observed in the blue arm (B arm) of the spectrograph, and only partially in the

red arm (R arm). DESI follows a linear wavelength grid with 0.8 Å steps.

DESI target selection is based on the public Legacy Surveys (Zou et al., 2017; Dey et al., 2019; Schlegel et al., 2023), with preliminary target selection details published for the MWS (Allende Prieto et al., 2020), BGS (Ruiz-Macias et al., 2020), LRGs (Zhou et al., 2020), ELGs (Raichoor et al., 2020) and quasars (Yèche et al., 2020) samples. A series of five papers describe the target selection for these objects (MWS, (Cooper et al., 2023); BGS, (Hahn et al., 2023); LRG, (Zhou et al., 2023); ELG, (Raichoor et al., 2023); quasar, (Chaussidon et al., 2023)).

## 4.1 Spectroscopic pipeline

While most of the fibers are assigned to science targets, some are used for calibration. Fibers assigned to the sky are essential for sky subtraction, as they enable the removal of the contribution to spectra caused by light from the background sky, particularly emission lines. Other fibers are assigned to standard stars in order to properly calibrate fluxes. This process is performed by comparing the measured counts in the CCD and the expected fluxes from the standard stars, allowing to estimate fluxes for all the other targets through a process called spectroperfectionism (described in Bolton and Schlegel (2010)). Spectroperfectionism provides spectral fluxes and their variances. For the full description of the DESI spectroscopic reduction pipeline see Guy et al. (2023).

Quasars at redshift higher than 2.1 are identified by the DESI pipeline as high redshift quasars suitable for Lyman- $\alpha$  forest studies. These will be re-observed, allowing for a better measurement of their spectra. Re-observations of the same quasar are then coadded and stored in files grouped by healpix pixels (see Górski et al. (2005)). The coaddition consists in a simple weighted average of fluxes for each wavelength in the spectra, using as weights their inverse-variance. Coadded files alongside the quasar catalog are used by our pipeline to build the Lyman- $\alpha$  forest catalog.

Each coadded file contains per-spectra information for the three arms (B, Z, R), as well as various metadata, including fiber positions, instrument configuration during observation and atmospheric conditions. The spectroscopic data include fluxes, estimated inverse-variance and a mask identifying invalid pixels.

## 4.2 Data Releases

DESI began the Survey Validation (SV) in December 2020, concluding it in April 2021. During this process, the performance of the instrument, the pipeline, and the data quality were tested.

The first data release of the collaboration consisted of these SV data, along with the commissioning data and some special observations (DESI Collaboration et al., 2023a). This was the Early Data Release (EDR). Several articles were published in the context of the Lyman- $\alpha$  forest analysis.

In the publication led by me (Ramírez-Pérez et al., 2024), we describe the catalog of Lyman- $\alpha$  fluctuations. This publication will be discussed in Chapter 6. The objective of Gordon et al. (2023) was to obtain the first Lyman- $\alpha$  correlation measurements from DESI early data, testing the pipeline and data quality, and compare its performance to previous eBOSS DR16 analyses. Herrera et al. (2023) provided details on the status of the different procedures used to build mocks for Lyman- $\alpha$  forest analyses. P1D analyses are performed in two different papers: Ravoux et al. (2023) presented the 1 dimensional measurement using Fast Fourier Transform (FFT), while Karaçaylı et al. (2024) made use of the Quadratic Maximum Likelihood Estimator (QMLE).

The DESI Data Release 1 (DR1) data release will be released at the end of 2024, BAO results from this dataset were published in April 2024, from galaxies and quasars (DESI Collaboration et al., 2024a) and from the Lyman- $\alpha$  forest (DESI Collaboration et al., 2024b). This sample covers data of the main survey starting in May 2021 until June 2022, with 12.8 million galaxies and quasars. In Chapter 7 we will show part of the validation performed for the DR1 Lyman- $\alpha$  forest BAO analysis.





# GENERATING SYNTHETIC DATASETS FOR COSMOLOGICAL PROBES WITH CoLoRe

The use of synthetic datasets is of vital importance in modern cosmological surveys. This is especially the case for the Lyman- $\alpha$  forest. With the current noise levels in spectrographs that can capture millions of objects, the measurement of the Lyman- $\alpha$  forest becomes highly complex, requiring precise validation of the tools used in the analysis. The CoLoRe package provides the main synthetic data used in the Lyman- $\alpha$  forest analysis for Dark Energy Spectroscopic Instrument (DESI).

In this Chapter, we present CoLoRe, a tool that efficiently generates synthetic data for multiple cosmological surveys, including Lyman- $\alpha$  forest, galaxy positions, lensing, integrated Sachs-Wolfe effect (ISW), and line intensity mapping. We also present the validation of the code and demonstrate its performance by simulating multiple surveys, including galaxies and quasars from DESI, galaxies and lensing from LSST, and SKA intensity mapping and radio galaxies.

The text from this Chapter is extracted from my publication Ramírez-Pérez et al. (2022), published in the Journal of Cosmology and Astroparticle Physics in May 2022.

This Chapter is organized as follows. In Section 5.1 we will provide more context to software for synthetic realisations and introduce the code. In Section 5.2 we describe in detail CoLoRe, its code structure, the cosmological assumptions made, and the list of tracers already available. The validation of CoLoRe to generate reliable mocks for intensity mapping and for Lyman- $\alpha$  forest studies was already presented in previous work (Alonso et al., 2014; Farr et al., 2020b). In Section 5.3 we validate its ability to simulate galaxy clustering and weak lensing statistics, and discuss the computing and memory requirements to run large simulations.

Finally, in Section 5.4 we draw some conclusions.

## 5.1 Introduction

Ongoing and future cosmological surveys will explore large volumes with multiple tracers to study Dark Energy (DE), inflation and massive neutrinos. Spectroscopic surveys such as the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al., 2016a), Euclid (Laureijs et al., 2011) and Roman (Spergel et al., 2015) will collect tens of millions of precise galaxy redshifts. Meanwhile, the Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al., 2009) will photometrically observe billions of galaxies and provide an exquisite weak lensing map over a large fraction of the sky. Lensing maps will also be obtained from future experiments observing the Cosmic Microwave Background (CMB), such as the Simons Observatory (SO) (Ade et al., 2019) and CMB-S4 (Abazajian et al., 2016). Finally, large catalogs of radio galaxies and 21cm intensity maps will be provided by experiments such as the Square Kilometer Array (SKA; Square Kilometre Array Cosmology Science Working Group et al., 2020) or HIRAX (Newburgh et al., 2016).

In order to obtain robust cosmological constraints from these large and complex datasets, it is important to be able to efficiently generate mocks, synthetic realisations of the data. For instance, mocks are often used to compare survey strategies, to test the analysis pipeline, to study the impact of astrophysical or instrumental contaminants, and to estimate the covariance of the measurements.

When generating mocks one needs to trade off realism for computing costs. It is just not feasible to generate hundreds or thousands of N-body simulations covering volumes of tens of cubic gigaparsecs, let alone hydrodynamic simulations that could model baryonic effects. On the other side of the spectrum, lognormal realisations (Coles and Jones, 1991) offer an efficient way of obtaining simplified mock catalogs with the correct distribution only on large, linear scales (see for instance Xavier et al. (2016), Agrawal et al. (2017), and Makiya et al. (2021)). Lagrangian Perturbation Theory (LPT) (Bernardeau et al., 2002) has inspired several approximated methods that can reproduce the distribution of matter on intermediate, mildly non-linear scales. These include PTHalos (Scoccimarro and Sheth, 2002; Manera et al., 2013, 2015), Pinocchio (Taffoni et al., 2002), COLA (Tassev et al., 2013), QPM (White et al., 2014), PATCHY (Kitaura et al., 2014) ICE-COLA (Izard et al., 2016), L-PICOLA (Howlett et al., 2015), HALOGEN (Avila et al., 2015) and EZMocks (Chuang et al., 2015).

The best cosmological inference will come from joint analyses of multiple cosmological probes, each providing independent and complementary information. Two of the most important challenges in these joint analyses will be characterising the effects of systematics affecting several experiments, and estimating the cross-covariance between the different 2D and 3D datasets, with partially overlapping area and redshift ranges. This present code addresses the need of simultaneously simulating these surveys in a coherent and efficient framework.

In this Chapter we present CoLoRe (Cosmological Lofty Realization), a parallel C code for generating fast mock realizations of multiple cosmological surveys<sup>1</sup>. CoLoRe can simulate the growth of structure using either a lognormal model or Lagrangian Perturbation Theory (LPT) (at 1st or 2nd order), and it can simulate a plethora of cosmological tracers: photometric and spectroscopic galaxies, weak lensing, intensity mapping, Integrated Sachs-Wolfe effect and CMB lensing, or the Lyman- $\alpha$  forest in the spectra of high-redshift quasars. It has been designed in a highly modular fashion, making it easy to add new tracers or more complex models of growth of structure. It uses both OpenMP and MPI parallelisation, and it is specially suited to run with multiple nodes in high performance computing facilities.

## 5.2 Methods

### 5.2.1 Overall code structure

CoLoRe is written in a modular way that makes it relatively straightforward to modify (e.g. to add a new non-linear structure formation model, or a new tracer of the density fluctuations). In a standard run, CoLoRe goes through the steps listed below, each of which is associated with a compartmentalised piece of code:

1. **Initialisation.** CoLoRe interprets the configuration file, allocates the resources needed to carry out the requested simulation, and initialises a number of cosmological quantities (redshift-distance relation, growth history, linear matter power spectrum, background densities of all source tracers).
2. **Predictions:** CoLoRe produces theoretical predictions for the three-dimensional power spectrum of all biased matter tracers in the lognormal approximation. This is mostly useful when using the lognormal structure formation model.

---

<sup>1</sup>The code is publicly available at <https://github.com/damonge/CoLoRe>.

3. **Gaussian random fields.** Two three-dimensional Cartesian grids are generated containing the linear matter overdensity  $\delta_M^L(\mathbf{x})$  and the Newtonian gravitational potential  $\phi_N(\mathbf{x})$  at redshift  $z = 0$ . The grid is sufficiently large to hold a sphere of comoving radius  $\chi(z_{\max})$ , where  $z_{\max}$  is the maximum redshift of the run. The spatial resolution of the simulation is set by  $N_{\text{grid}}$ , the number of grid cells into which the box is divided in each dimension. The grid cell size is therefore approximately  $\Delta x = L_{\text{box}}/N_{\text{grid}} \simeq 2\chi(z_{\max})/N_{\text{grid}}$ .
4. **Physical density field.** The Gaussian overdensity  $\delta_M^L(\mathbf{x})$  is transformed into a non-linear, physical overdensity field  $\delta_M(\mathbf{x})$  through one of the structure formation models supported by CoLoRe. This is done in the lightcone (i.e. the value of the field at comoving position  $\mathbf{x}$  is  $\delta_M(t(|\mathbf{x}|), \mathbf{x})$ , where  $t(\chi)$  is the cosmic time at comoving distance  $\chi$ ), with the observer located at the center of the Cartesian box. The physical density field is such that  $\delta \geq -1$  everywhere. The gravitational potential is also evolved in the lightcone assuming linear growth.
5. **Density normalisation.** CoLoRe uses non-linear transformations to generate biased tracers of the matter overdensity. In general, these can be written as

$$1 + \delta_k = \frac{B_k(\delta_M)}{\langle B_k(\delta_M) \rangle}, \quad (5.1)$$

where  $B_k$  is the non-linear biasing relation for tracer  $k$ . Due to the non-linearity of these relations, the ergodic average in the denominator of the previous equation is not necessarily equal to 1 (even if  $\langle \delta_M \rangle = 0$ ), and therefore the normalising factor ensures that  $\langle \delta_k \rangle = 0$  for all biased tracers. Since densities are defined on the lightcone, the normalisation factors are computed at this stage independently for several redshift shells by averaging over grid cells.

6. **Get Cartesian information.** At this stage the overdensity and Newtonian potential grids are distributed across computer nodes as slabs of equal width  $N_{\text{slab}} = N_{\text{grid}}/N_{\text{nodes}}$ . Before proceeding further, CoLoRe collects all information available in these slabs and needed by each of the tracers requested for this simulation. This involves any data product not involving line-of-sight integrals or interpolations, which are dealt with in later stages. For instance, this is when source catalogs are generated by Poisson-sampling the biased density field. All tracers are endowed with a method `tracer_set_cartesian` that collects this information.

7. **Redistribution into beams.** In order to carry out line-of-sight integrals and interpolations, CoLoRe redistributes tracer data so each node has access to all the data in a set of sky regions, labelled “beams”, covering the full range of redshifts  $0 \leq z \leq z_{\max}$ . Each beam is defined using the HEALPix pixellation scheme (Górski et al., 2005), as the region of the celestial sphere covered by a given low resolution pixel. The HEALPix  $N_{\text{side}}$  resolution parameter used to define these beams is chosen to be large enough that the full dataset is approximately evenly split between computer nodes. All tracers have an associated method `tracer_distribute` in charge of distributing the tracer data in each node’s slab to all other nodes whose beams intersect with it. Note that, although the tracer information is now distributed across nodes through beams, the density and Newtonian potential grids are still distributed in slabs.
8. **Get beam information.** Any calculation involving a line-of-sight integral (e.g. gravitational lensing) or interpolation (e.g. Lyman- $\alpha$  skewers) is done after the tracers have been redistributed into beams. The calculation is done in three stages:
  - a) *Preprocessing.* Initialisation of any necessary quantities (e.g. setting all variables that eventually hold a gravitational lensing calculation to zero). Each tracer has an associated function `tracer_beams_preproc` in charge of doing this.
  - b) *Loop through slabs.* All nodes send their current slab of the density and Newtonian potential grids to the node on their right (assuming periodic boundary conditions). Once the new slab is received, each node gathers the necessary information from it (e.g. the contribution to a lensing convergence integral from the Section of the Newtonian potential held in that slab) and adds it to each tracer<sup>2</sup>. CoLoRe carries out this calculation through a method `tracer_get_beam_properties` associated with each tracer. This is repeated  $N_{\text{nodes}}$  times, at which point all nodes have had access to the full density and potential grids.
  - c) *Post-processing.* Each tracer finishes off any calculation still needed after having gathered all information in the preceding step (e.g. multiplying maps by an overall normalization factor). This is done by a method `tracer_beams_postproc` associated with each tracer.

---

<sup>2</sup>Note that all quantities calculated at this stage (e.g. integrals and interpolations) are linear and additive on  $\delta_M$  and  $\phi_N$ .

9. **Write output.** Each tracer writes all its information (e.g. in the form of maps or catalogs) to file through a method `write_tracer`. CoLoRe uses the FITS standard in most cases, although it is also possible to save source catalog data as ASCII or HDF5 files.

Modifying CoLoRe to support a new structure formation model would involve implementing it as part of step 4 above, with no effect on the rest of the code. Adding a new type of tracer would involve creating the corresponding tracer methods for it enumerated above (`_set_cartesian`, `_distribute`, `_beams_preproc`, `_get_beam_properties`, `_beams_postproc`, and `write_`).

The assumptions made by CoLoRe to compute the background cosmological quantities (step 1) are described in Section 5.2.2. Section 5.2.3 describes the Gaussian density fields and the different non-linear structure formation models supported by CoLoRe (steps 3 and 4). Section 5.2.4 describes in detail the calculations carried out for each of the tracers, including the bias models supported (steps 5-8). Finally, the theory predictions computed by CoLoRe for lognormal fields (step 2) are discussed in Section 5.2.5.

## 5.2.2 Cosmological assumptions

When generating simulated observations, CoLoRe makes a number of assumptions about the underlying cosmological model. We describe these here.

CoLoRe assumes a flat  $w$ CDM cosmological background, characterised, at low redshifts, by 3 cosmological parameters: the background matter density  $\Omega_M$ , the current expansion rate  $H_0$ , and a constant DE equation of state parameter  $w$ . The expansion rate is thus given by

$$H(z) = H_0 \left[ \Omega_M(1+z)^3 + (1 - \Omega_M)(1+z)^{3(1+w)} \right], \quad (5.2)$$

in terms of which the comoving distance is<sup>3</sup>

$$\chi(z) = \int_0^z \frac{dz'}{H(z')}. \quad (5.3)$$

Matter density perturbations are governed by a linear matter power spectrum at  $z = 0$ ,  $P_0(k)$ , which must be provided to CoLoRe on input, and is then normalised to the chosen value of  $\sigma_8$ . If the power spectrum is needed on scales larger than those provided, it is extrapolated assuming a power-law behaviour  $P_0(k) \propto k^{n_s}$  on small  $k$ , where  $n_s$  is the scalar spectral index (also provided on input).

---

<sup>3</sup>Note that we use units with  $c = 1$  throughout.

Finally, CoLoRe assumes a self-similar growth for the linear matter overdensity:  $\delta_M^L(\mathbf{x}, z) = \delta_M^L(\mathbf{x}, 0)D(z)$ , where  $D(z)$  is the linear growth factor.  $D(z)$  is calculated from the cosmological parameters by solving the differential equation

$$\frac{d}{da} \left( a^3 H(a) \frac{dD}{da} \right) = \frac{3}{2} \Omega_M(a) a H(a) D(a), \quad (5.4)$$

where  $a = 1/(1+z)$  is the scale factor.

Although internally CoLoRe uses “ $h$ -inverse” units (i.e. distances are given in units of  $\text{Mpc } h^{-1}$ ), all simulation outputs involve observable quantities (redshift and angles), and therefore are insensitive to this choice.

### 5.2.3 Matter box

The first step after initialising the cosmological model in a standard CoLoRe run is the generation of a Gaussian realisation of the linear matter inhomogeneities  $\delta_M^L$  at  $z = 0$  on a Cartesian cubic grid. This is done by drawing the Fourier coefficients of  $\delta_M^L$  as independent Gaussian random numbers from the input linear matter power spectrum using the Box-Muller transform with variance:

$$\sigma^2(\mathbf{k}) = \frac{P_0(k)}{(\Delta k)^3}, \quad (5.5)$$

where  $\Delta k \equiv 2\pi/L_{\text{box}}$  is the sampling rate in Fourier space. CoLoRe can alternatively apply a Gaussian smoothing kernel with scale  $R_G$  to the linear power spectrum when generating the linear Fourier coefficients. This may be useful to control the behaviour of the non-linear overdensity field (see discussion in Section 5.2.5).

At the same time, CoLoRe populates a similar cartesian grid with the values of the Newtonian gravitational potential  $\phi_N(\mathbf{x})$ , related to the matter inhomogeneities in Fourier space via:

$$\phi_N(\mathbf{k}) = -\frac{3}{2} H_0^2 \Omega_M \frac{\delta_M^L(\mathbf{k})}{k^2}. \quad (5.6)$$

The linear matter overdensity thus generated is then transformed into a physical (i.e. positive-definite) non-linear matter overdensity in the lightcone using one of the three structure formation models currently supported by CoLoRe, which we describe below.

#### 5.2.3.1 Lognormal fields

Lognormal fields were first proposed and analysed by Coles and Jones (1991) as a possible way to describe the distribution of matter in the universe. A lognormal



random field  $x_{\text{LN}}$  is defined in terms of a Gaussian random field  $x_{\text{G}}$  through the local transformation

$$x_{\text{LN}} = \exp x_{\text{G}}. \quad (5.7)$$

One of the nice properties of these fields is that, while the Gaussian variable  $x_{\text{G}}$  is allowed to take any values in  $(-\infty, +\infty)$ ,  $x_{\text{LN}}$  can only take positive values by construction. Furthermore, as discussed in Coles and Jones (1991), the density field evolved along Lagrangian trajectories according to the linear velocity field along can be well described by a lognormal distribution, which justifies the use of lognormal fields from a physical point of view. In order to obtain a lognormal overdensity field with zero mean from a Gaussian field, the transformation (Eq. (5.7)) must be slightly varied as follows:

$$1 + \delta_{\text{LN}} = \exp \left( \delta_{\text{G}} - \frac{\sigma_{\text{G}}^2}{2} \right), \quad (5.8)$$

where  $\sigma_{\text{G}}$  is the variance of the Gaussian overdensity field.

Lognormal density fields have been used in the past by different collaborations in order to perform fast galaxy mock realisations (Cole et al., 2005; Beutler et al., 2011; Blake et al., 2011; Le Goff et al., 2011; Font-Ribera et al., 2012a), and are, therefore, a well established tool. Since the lognormal transformation is a simple, local modification of the density field, it is by far the fastest and most memory-efficient structure formation model implemented in CoLoRe.

However, the simplicity of the lognormal transformation implies that lognormal fields cannot be expected to describe all higher-order correlators of the density field (e.g. bispectra), to give rise to filamentary structure, or to reproduce the small-scale properties of the density field correctly (Kitaura et al., 2010), and therefore this kind of mock realisations have a limited applicability.

Within this framework, the non-linear overdensity is generated in CoLoRe by applying Eq. (5.8) to the linear overdensity field evolved in the lightcone assuming linear growth (which is applied to both  $\delta_{\text{G}}$  and  $\sigma_{\text{G}}$  in this equation).

Caution must be exercised when making use of lognormal fields. Since the lognormal transformation involves the exponentiation of a Gaussian field, large ( $\gg 1$ ) values of  $\delta_{\text{G}}$  lead to much larger fluctuations in  $\delta_{\text{LN}}$ . Thus, if the amplitude of  $\delta_{\text{G}}$ , characterised by its standard deviation  $\sigma_{\text{G}}$ , is large, the resulting lognormal field will exhibit a large degree of inhomogeneity, with an enormous variance dominated by extreme fluctuations in a small number of voxels. This behaviour can be avoided through the use of the Gaussian smoothing kernel described above. This modifies the linear and non-linear power spectra in an analytically predictable manner.

Note that other common implementations of the lognormal transformation to generate mock cosmological realisations (e.g. Beutler et al. (2011)) have employed a different method to avoid this problem. Instead of using the input power spectrum to generate  $\delta_G$  and then transform it into  $\delta_{LN}$ , the input power spectrum is taken to be that of the final  $\delta_{LN}$ . The inverse lognormal transformation is thus applied to the input power spectrum at the level of the two-point correlator (see Section 5.2.5), to obtain the power spectrum with which the Gaussian field is generated. CoLoRe does not explicitly support this method, since it runs contrary to the idea of treating the lognormal transformation as a non-linear structure formation model. It would be, however, possible to pass as input to CoLoRe the “Gaussianised” power spectrum in order to obtain the desired matter power spectrum at  $z = 0$  in the non-linear matter fluctuations.

### 5.2.3.2 Lagrangian perturbation theory

Lagrangian Perturbation Theory (LPT) (Bernardeau et al., 2002) provides an alternative fast method to generate non-linear physical matter overdensities, which has been used in the past to generate mock galaxy catalogs (Manera et al., 2013; Chuang et al., 2015; Manera et al., 2015). CoLoRe supports the generation of Lagrangian displacements at first and second order in perturbation theory. These first and second-order displacements at  $z = 0$  are scaled with the corresponding growth factors before interpolating the test particles onto a grid. We provide a brief overview of LPT here, and direct the reader to (Bernardeau et al., 2002) for further details.

Let  $\mathbf{x}(t) = a(t) [\mathbf{q} + \mathbf{\Psi}(\mathbf{q}, t)]$  be the physical position of a particle starting at comoving coordinates  $\mathbf{q}$ .  $\mathbf{\Psi}(\mathbf{q}, t)$  is the so-called Lagrangian displacement vector. In the Newtonian approximation, the motion of these particles is governed by Newton’s second law, which is sourced by the gravitational potential caused by the particles themselves. This leads to two coupled equations that can be summarized into a single equation for the divergence of  $\mathbf{\Psi}$ :

$$J \left( J^{-1} \nabla \right) \cdot (\mathbf{\Psi}'' + a H \mathbf{\Psi}') = \frac{3}{2} a^2 H^2 \Omega_M (J - 1), \quad (5.9)$$

where all derivatives are taken with respect to conformal time  $d\tau \equiv dt/a$ ,  $J_{ij} \equiv \delta_{ij} + \partial_i \Psi_j$  is the Jacobian of the Lagrangian flow, and  $J \equiv \det(J)$ . Note that the matter overdensity is given by  $1 + \delta = J^{-1}$ .

At second order in the displacement field, and discarding all curl-like components of  $\mathbf{\Psi}$ , the solution is given by

$$\mathbf{\Psi}(\mathbf{q}, a) = D(a) \nabla \varphi_{\text{LPT}}^{(1)}(\mathbf{q}) + D^{(2)}(a) \nabla \varphi_{\text{LPT}}^{(2)}(\mathbf{q}). \quad (5.10)$$

Here  $D(a)$  is the linear growth factor, satisfying Eq. (5.4),  $D^{(2)}$  is the second-order growth factor, satisfying

$$\frac{d}{da} \left( a^3 H(a) \frac{dD^{(2)}}{da} \right) = \frac{3}{2} \Omega_M(a) a H(a) \left( D^{(2)}(a) - [D(a)]^2 \right), \quad (5.11)$$

and  $\varphi_{\text{LPT}}^{(1,2)}$  are the first- and second-order LPT potentials, given by

$$\nabla^2 \varphi_{\text{LPT}}^{(1)} = -\delta_M^L, \quad (5.12)$$

$$\nabla^2 \varphi_{\text{LPT}}^{(2)} = \frac{1}{2} \sum_{ij} \left[ \partial_i^2 \varphi_{\text{LPT}}^{(1)} \partial_j^2 \varphi_{\text{LPT}}^{(1)} - \left( \partial_i \partial_j \varphi_{\text{LPT}}^{(1)} \right)^2 \right]. \quad (5.13)$$

CoLoRe solves the LPT Poisson equations (Eq. (5.12) and Eq. (5.13)) in Fourier space using the Gaussian overdensity field as input, and computes the second-order growth factor using the approximation (Bernardeau et al., 2002)

$$D_2(a) = -\frac{3}{7} [D(a)]^2 [\Omega_M(a)]^{-1/143}. \quad (5.14)$$

Once the displacement vector has been calculated and applied to a set of test particles initially located at the centers of the Cartesian grid cells, the density field is calculated by interpolating the displaced positions onto the grid<sup>4</sup>.

LPT is able to generate a more realistic non-linear density field than the lognormal model at the cost of significantly higher memory requirements and longer computation times. While generating the lognormal overdensity does not require additional resources beyond those used to generate the  $\delta_M^L$  and  $\phi_N$  grids, generating the first-order displacement requires three additional Cartesian grids to hold the components of  $\Psi$ , and second-order LPT demands an additional 5 grids to store the Hessian of the first-order LPT potential (the array holding the Gaussian density field can be reused to store one of the 6 independent components of the Hessian). The number of fast Fourier transforms needed, which dominate the total computation time, is also different in each case: none in the case of the lognormal model, 4 in the case of first-order LPT, and 13 for second-order LPT.

## 5.2.4 Tracers

CoLoRe is able to generate simulated observations for a variety of cosmological tracers of the same underlying matter fluctuations. The details of the calculations involved in each of the supported tracer types, carried out in steps 5 to 8 of the procedure outlined in Section 5.2.1, is discussed in detail here.

---

<sup>4</sup>To do this, CoLoRe supports three standard mass-assignment methods: nearest-grid-point (NGP), cloud-in-cell (CIC), and triangular-shaped cloud (TSC) (Hockney and Eastwood, 1981).

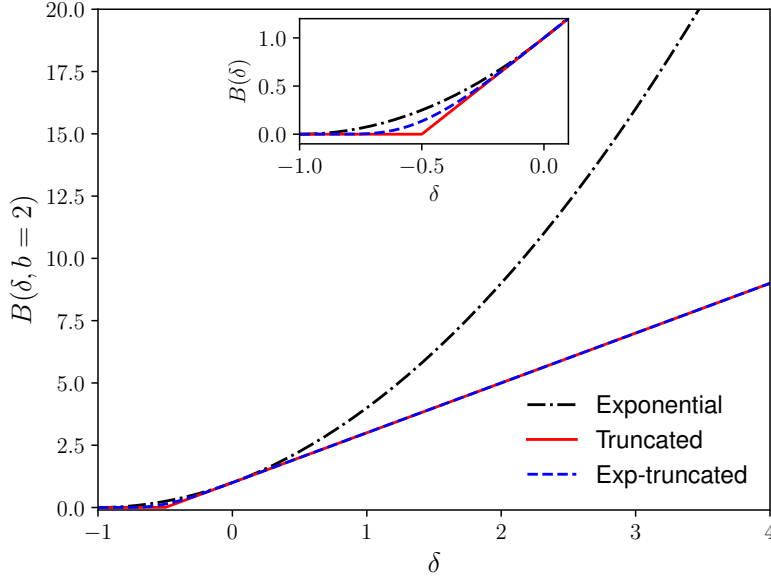


Figure 5.1: Bias models implemented in CoLoRe. The exponential model preserves the “lognormality” of the field if using the lognormal structure formation model, but it can lead to numerically unstable results in the presence of sufficiently large fluctuations in the Gaussian field. The exp-truncated model can be used to curb this behaviour.

#### 5.2.4.1 Projected maps

The simplest tracer supported by CoLoRe is the so-called “custom projected tracer”. The associated observable is the overdensity in a biased tracer of the matter fluctuations projected onto the celestial sphere after integrating over an arbitrary radial kernel:

$$\Delta_W(\hat{\mathbf{n}}) = \int dz W(z) \delta_W(z, \chi(z)\hat{\mathbf{n}}), \quad (5.15)$$

where the integral is over redshift  $z$ ,  $W(z)$  is the tracer’s radial kernel, and  $\delta_W$  are the three-dimensional fluctuations in the tracer, related to the matter fluctuations via Eq. (5.1).

CoLoRe supports the following three local bias models, although more can be easily added to the code:

$$\textbf{Exponential:} \quad B_k(\delta_M) = (1 + \delta_M)^{b_k}, \quad (5.16)$$

$$\textbf{Truncated:} \quad B_k(\delta_M) = \text{Max}(1 + b_k \delta_M, 0), \quad (5.17)$$

$$\textbf{Exp-truncated:} \quad B_k(\delta_M) = \begin{cases} \exp[b_k \delta_M / (1 + \delta_M)] & \delta_M \leq 0 \\ 1 + b_k \delta_M & \delta_M > 0 \end{cases}. \quad (5.18)$$

These three models are shown in Fig. 5.1 for  $b_k = 2$ , and are designed to be positive definite ( $B_k(\delta_M) > 0$  for  $\delta_M \in [-1, \infty)$ ), and to reduce to a linear biasing relation

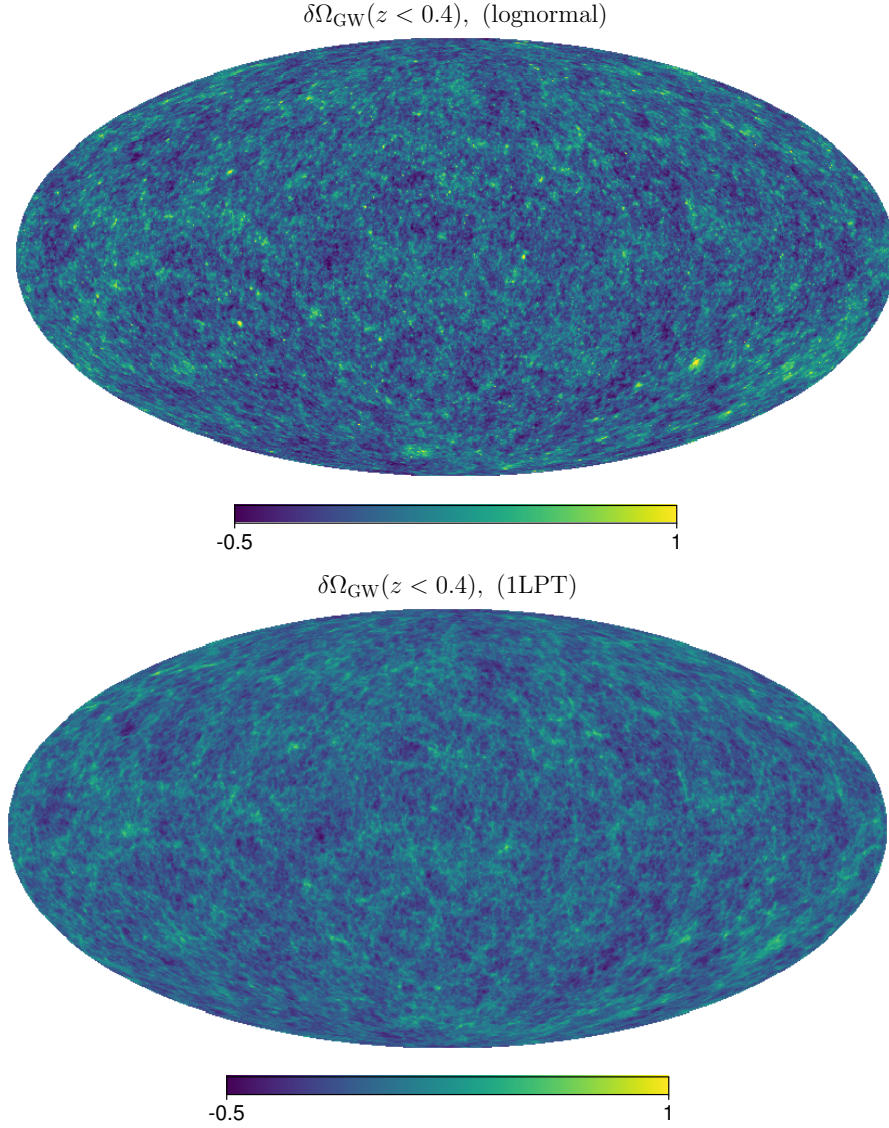


Figure 5.2: Simulated maps of the anisotropic stochastic gravitational wave background from astrophysical sources at redshifts  $z < 0.4$  using the models of (Cusin et al., 2020). The top and bottom plots show simulations using the lognormal and first-order LPT structure formation models respectively. The former is characterised by strong positive fluctuations on a few regions, while the latter displays the more physical filamentary structure of the cosmic web.

$(B_k(\delta_M) \simeq 1 + b_k \delta_M)$  for small fluctuations. Given the modular nature of CoLoRe, different biasing models could be easily added by the user.

The integral in Eq. (5.15) is calculated as follows: for each sky pixel, an imaginary line of sight connecting it with the observer at the box center is subdivided into intervals of constant comoving distance  $\Delta\chi$ , commensurate with the Cartesian cell size  $\Delta x$ . The value of  $\delta_M$  at the center of each interval is calculated from the Cartesian grid using trilinear interpolation, and is then translated into the corresponding  $\delta_W$ . The integral is then calculated as a sum over all intervals along the line of sight.

Custom projected tracers can be used in CoLoRe to make simulated maps of a wide variety of two-dimensional cosmological anisotropic observables that correlate with the large scale structure (LSS). As an example, Fig. 5.2 shows simulated maps of the anisotropic stochastic gravitational wave background from astrophysical sources at  $z < 0.4$  according to the models of Cusin et al. (2020). The figure shows results for the lognormal and first-order LPT structure formation models, showcasing the morphological differences between them.

#### 5.2.4.2 Integrated Sachs-Wolfe effect

The time evolution in the gravitational potential at late times due to the accelerated background expansion causes an energy loss or gain in a background of photons which correlates with the LSS. This is the so-called integrated Sachs-Wolfe effect (ISW; Sachs and Wolfe, 1967).

The fluctuation in the temperature of a background of photons with black-body spectrum emitted at redshift  $z_*$  is given by

$$\left. \frac{\Delta T}{T} \right|_{\text{ISW}}(\hat{\mathbf{n}}) = 2 \int_0^{z_*} dz \frac{\dot{\phi}_N(z, \chi(z)\hat{\mathbf{n}})}{(1+z)H(z)}. \quad (5.19)$$

Assuming linear growth, appropriate on the large scales on which the ISW is relevant, one can approximate  $\dot{\phi}_N(z) = H(z)[f(z)-1]\phi_N(z)$ , where  $f \equiv d \log D / d \log a$  is the growth rate.

The ISW tracer is therefore equivalent to the custom projected tracer described in the previous Section with the Newtonian potential  $\phi_N$  taking the role of  $\delta_W$ , and with a kernel

$$W_{\text{ISW}}(z) = \frac{f(z)-1}{1+z} \Theta(z < z_*), \quad (5.20)$$

where  $\Theta$  is the Heaviside function. Thus, the same numerical methods described in Section 5.2.4.1 are used by CoLoRe to generate simulated ISW maps. The top panel of Fig. 5.3 shows an example of a simulated map of the ISW effect for  $z_* = 0.5$ ,



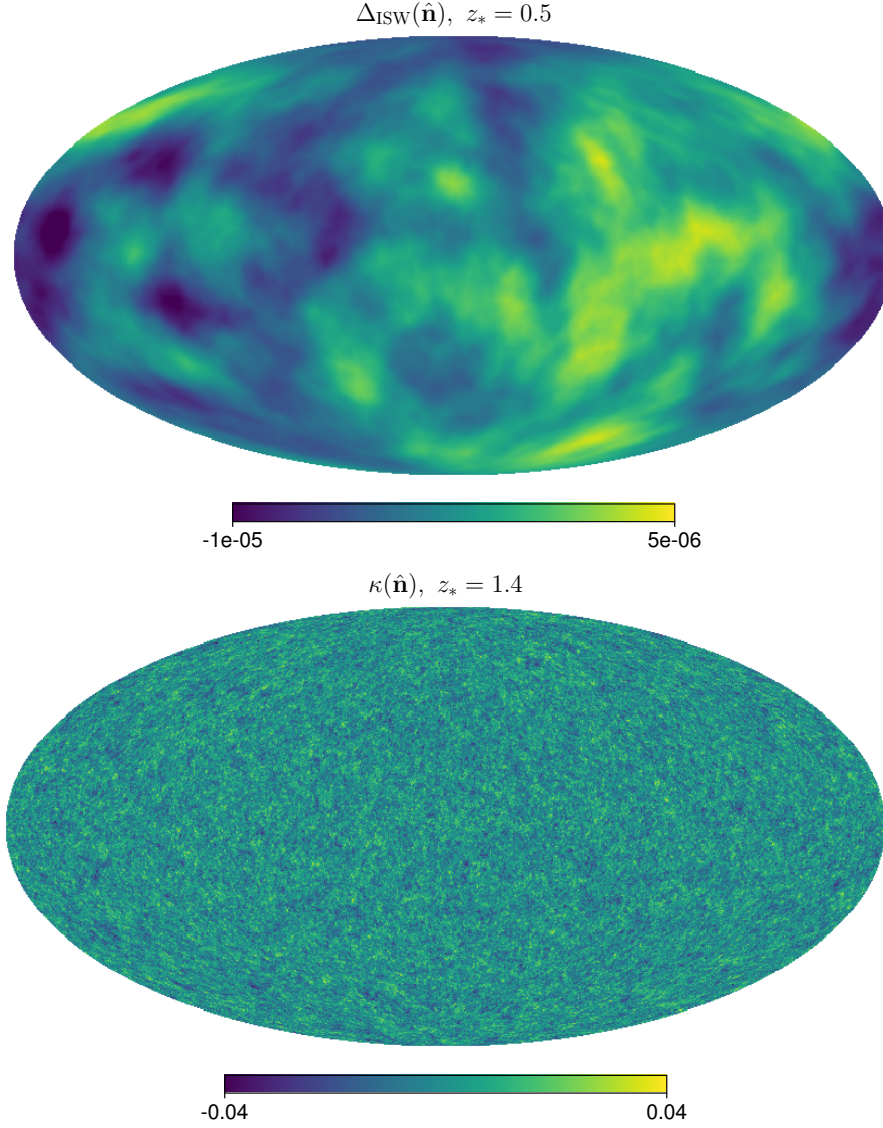


Figure 5.3: *Top*: simulated map of the low-redshift ISW effect for a source plane at  $z_* = 0.5$ . *Bottom*: map of the lensing convergence for a source plane at  $z_* = 1.4$ .

characterised by features on very large scales due to the  $1/k^2$  relation between gravitational potential and matter overdensity.

#### 5.2.4.3 Gravitational lensing

As we saw in Section 3.4, the fluctuations in the gravitational potential perturb the trajectories of photons. Computing  $\alpha \equiv \nabla_{\hat{\mathbf{n}}} \psi$  and  $\kappa$  (defined in Eq. (3.24)) from the lensing potential  $\psi$  (Eq. (3.23)) would require first creating a map of  $\psi$  to then differentiate. This is not feasible when calculating the effects of lensing on a large number of sources at different redshifts (see Section 5.2.4.4). Instead, we rewrite

these quantities as

$$\alpha(\hat{\mathbf{n}}) = -2 \int_0^{\chi_*} d\chi \frac{\chi_* - \chi}{\chi_*} \nabla_{\perp} \phi_N(z, \chi \hat{\mathbf{n}}), \quad \gamma(\hat{\mathbf{n}}) = -2 \int_0^{\chi_*} d\chi \chi \frac{\chi_* - \chi}{\chi_*} \mathbf{H}_{\perp} \phi_N(z, \chi \hat{\mathbf{n}}), \quad (5.21)$$

where  $\nabla_{\perp}$  and  $\mathbf{H}_{\perp}$  are the gradient and Hessian operators projected onto the plane perpendicular to  $\hat{\mathbf{n}}$ .

Explicitly, if  $\hat{\mathbf{n}} \equiv (\sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta)$ , defining the projector

$$\mathbf{P}_{\hat{\mathbf{n}}} \equiv \begin{pmatrix} \cos \theta \cos \varphi & \cos \theta \sin \varphi & -\sin \theta \\ -\sin \varphi & \cos \varphi & 0 \end{pmatrix}, \quad (5.22)$$

the projected gradient and Hessian are

$$\nabla_{\perp} \phi_N \equiv \mathbf{P}_{\hat{\mathbf{n}}} \nabla \phi_N, \quad \mathbf{H}_{\perp} \phi = \mathbf{P}_{\hat{\mathbf{n}}} (\mathbf{H} \phi_N) \mathbf{P}_{\hat{\mathbf{n}}}^T, \quad (5.23)$$

where  $\nabla_i \equiv \partial/\partial x^i$  and  $\mathbf{H}_{ij} \equiv \partial^2/\partial x^i \partial x^j$ .

The calculation of lensing-related quantities in CoLoRe is thus analogous to the procedure outlined in Section 5.2.4.1: along a given line of sight  $\hat{\mathbf{n}}$  (corresponding to a map pixel or to the position of a given source), the values of the first or second-order derivatives of  $\phi_N$  are calculated and interpolated from the Cartesian grid onto a set of equidistant points along  $\hat{\mathbf{n}}$ . The corresponding quantities are then projected onto the plane perpendicular to  $\hat{\mathbf{n}}$ , and the integrals in Eq. (5.21) are computed as direct sums over the evaluated points.

Besides providing lensing information associated with its source catalogs (see Section 5.2.4.4), CoLoRe returns maps of the lensing convergence  $\kappa(\hat{\mathbf{n}})$  for an arbitrary number of source planes at different redshifts. An example at  $z_* = 1.4$  is shown in the bottom panel of Fig. 5.3.

#### 5.2.4.4 Sources

CoLoRe can also generate catalogs of discrete sources as biased tracers of the matter distribution. The source distribution is modelled as a Cox process: the number of sources of type  $a$  in a given Cartesian cell  $i$ ,  $N_i^a$  is a random Poisson variable with a stochastic mean given by

$$\bar{N}_i^a = (\Delta x)^3 \bar{n}_a(\chi) \frac{B_a(\delta_M(\mathbf{x}_i))}{\langle B_a(\delta_M) \rangle}, \quad (5.24)$$

where  $\mathbf{x}_i$  are the coordinates of cell  $i$ ,  $(\Delta x)^3$  is the cell volume,  $\bar{n}_a(\chi)$  is the redshift-dependent mean density of sources, and  $B_a$  is the biasing relation of type- $a$  sources (see Eqs. (5.16) to (5.18)).



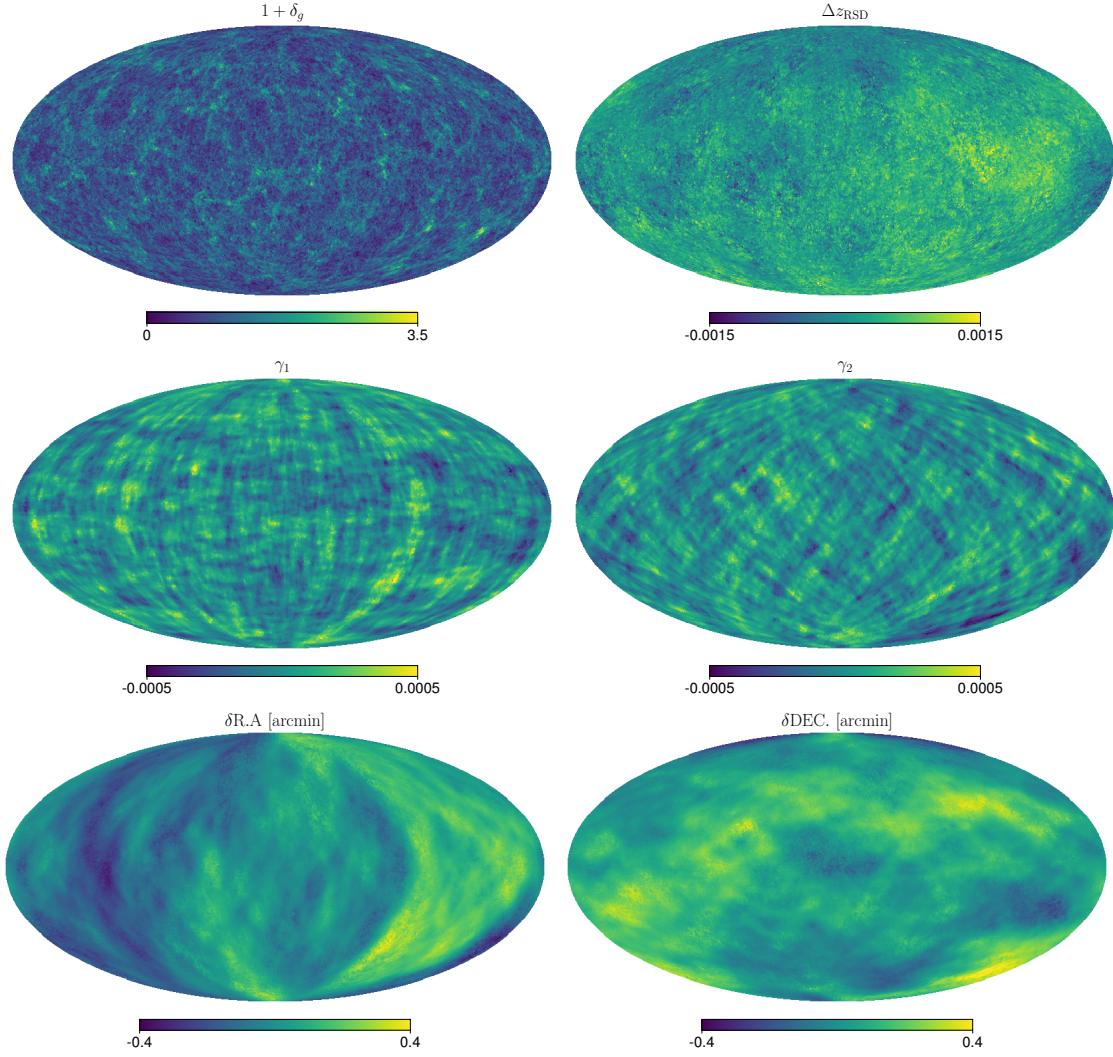


Figure 5.4: Maps of source-related quantities for a simulated CoLoRe catalog in the redshift range  $z < 0.3$ . *Top left*: source overdensity. *Top right*: mean redshift distortion. *Middle*: mean lensing shear. *Bottom*: mean lensing displacement vector.

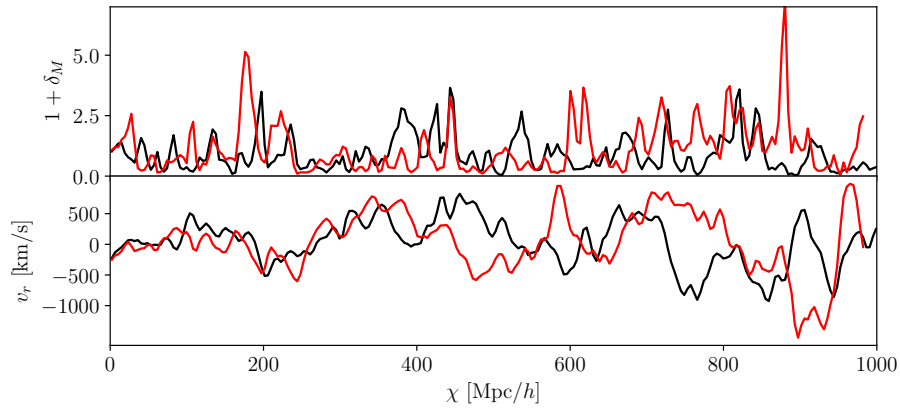


Figure 5.5: Density (*top*) and radial velocity (*bottom*) skewers for two arbitrary sources in a simulated CoLoRe catalog.

Once  $N_i^a$  has been determined for each cell, the corresponding number of sources are generated with three-dimensional positions randomly located within their cell. The three-dimensional comoving position is then translated into observable angular and redshift coordinates. In addition to the cosmological redshift, the contribution from the source’s peculiar velocity, necessary to simulate redshift-space distortions, is calculated as  $\Delta z_{\text{RSD}} = v_r$ , where the radial velocity is calculated from the gradient of the gravitational potential as

$$v_r(z, \mathbf{x}) = -\frac{2}{3H_0^2\Omega_M}f(z)(\hat{\mathbf{n}} \cdot \nabla)\phi_N(z, \mathbf{x}). \quad (5.25)$$

Given this derivation of velocities they will only include linear effects.

If desired, the source catalogs can also contain gravitational lensing information  $(\alpha_\theta, \alpha_\varphi, \kappa, \gamma_1, \gamma_2)$  for each source. This can be used to construct a weak lensing shear catalog with ellipticities  $e_i = \gamma_i$ , or to include the effects of lensing magnifications by perturbing the source angular positions  $\theta \rightarrow \theta + \alpha_\theta$ ,  $\varphi \rightarrow \varphi + \alpha_\varphi$ , and its flux by  $F \rightarrow F(1 + 2\kappa)$ . In this case, all lensing quantities are calculated as described in Section 5.2.4.3, by integrating the interpolated transverse derivatives of the gravitational potential along each source’s line of sight.

Source catalogs can also be endowed with so-called “line-of-sight skewers”, containing the matter overdensity and radial velocity interpolated from the Cartesian box onto each source’s line of sight. As in the case of gravitational lensing calculations, trilinear interpolation is used, and both fields are sampled at radial comoving intervals equal to CoLoRe’s Cartesian cell size. The use of skewers to produce simulated observations of the Lyman- $\alpha$  forest was discussed in detail in (Farr et al., 2020b).

It is worth noting that the computation of any quantity requiring full line-of-sight information (lensing or skewers) requires redistributing sources from “slabs” into “beams” across nodes, and for all nodes to loop through the full Cartesian box to add the contribution of all slabs to their beams (i.e. points 7 and 8 in Section 5.2.1). This process requires significant communication between MPI nodes and, depending on the particular case, can have a significant impact on the total computing time. For this reason, if no line-of-sight information is requested from CoLoRe (i.e. if only sources or intensity maps are requested, with no associated lensing information or skewers), steps 7 and 8 are skipped, often leading to a significant speed-up.

Furthermore, computing the lensing observables along each line-of-sight for large catalogs containing billions of sources, as would be the case e.g. if simulating the full LSST shear sample, is a computationally demanding task that can dominate

by far all other steps in a CoLoRe run. To avoid this, an alternative approximate scheme to compute these quantities can be used. In the “fast lensing” scheme, the five lensing observables (displacement, convergence and shear) are precomputed on a set of spherical HEALPix maps at constant radial comoving distance intervals, using the same method outlined in Section 5.2.4.3. The angular resolution parameter  $N_{\text{side}}$  of each map is adaptively chosen so that the physical pixel size is smaller than the Cartesian cell size, in order to avoid oversampling as well as degrading the original three-dimensional resolution. The hierarchical nature of the HEALPix scheme makes it possible to relate a pixel in a given map with pixels in all lower-resolution maps with smaller radii. The lensing observables for a given source are then calculated by interpolating between the values of that observable along the line of pixels corresponding to the source’s angular coordinates.

Fig. 5.4 shows maps of the quantities described in this Section for a small source catalog simulated with CoLoRe. The simulated sample had a redshift distribution

$$\frac{dN}{dz} \propto \left(\frac{z}{z_0}\right)^2 \exp\left[-\left(\frac{z}{z_0}\right)^{3/2}\right], \quad (5.26)$$

with  $z_0 = 0.07$ , extending up to  $z \lesssim 0.3$ . The simulation was run using the first-order LPT structure formation model. The different panels show the source overdensity and mean redshift distortion (top panels), the mean lensing shear (middle panels) and the mean lensing deflection vector (bottom panels). Fig. 5.5 shows density and velocity skewers (top and bottom panels respectively) for two arbitrary sources in the same catalog.

#### 5.2.4.5 Line intensity mapping

Consider a species of gas emitting at a rest-frame frequency  $\nu_0$  due to some atomic or molecular transition. The intensity (flux per unit frequency) measured in a patch around  $\hat{\mathbf{n}}$  with solid angle  $\delta\Omega$ , and in a frequency interval  $\delta\nu$  around  $\nu$  is given by (Abdalla and Rawlings, 2005)

$$I(\nu, \hat{\mathbf{n}}) = \frac{\hbar A_{21} \nu_0 x_2}{2m_a} \frac{M_{\text{em}}}{(1+z)^2 \chi^2 \delta\Omega \delta\nu}, \quad (5.27)$$

where  $A_{21}$  is the Einstein coefficient for the transition,  $m_a$  is the atomic mass of the emitting gas,  $x_2$  is the fraction of the gas in the excited state,  $M_{\text{em}}$  is the total mass of the emitting gas in the voxel defined by  $(\delta\nu, \delta\Omega)$ . Associating this intensity with a black-body temperature in the Rayleigh-Jeans regime ( $T = c^2 I / (2k_B \nu^2)$ ), this can be rewritten as:

$$T(\nu, \hat{\mathbf{n}}) = \bar{T}(z) [1 + \delta_{\text{em}}(z, \hat{\mathbf{n}})], \quad (5.28)$$

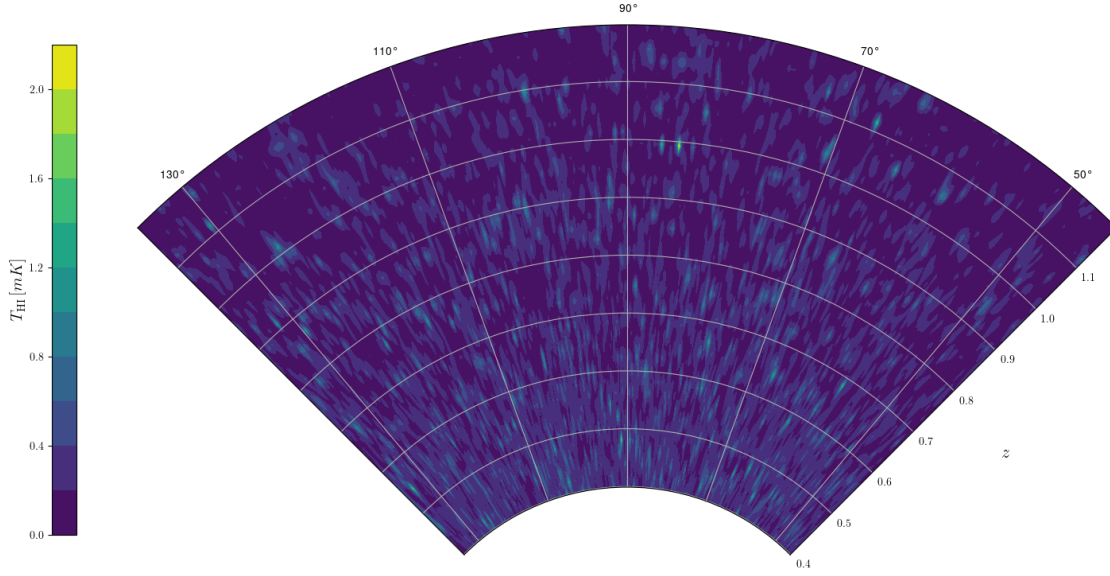


Figure 5.6: Slice through one of the 21cm intensity maps generated with CoLoRe.

where  $\delta_{\text{em}}$  is the overdensity of the emitting gas in redshift space (i.e. accounting for the effects of redshift-space distortion),  $z = \nu_0/\nu - 1$ , and the mean temperature  $\bar{T}$  is

$$\bar{T}(z) \equiv \frac{3 \hbar A_{21} x_2 x_{\text{em}}(z) \Omega_{b,0} H_0^2 c^2 (1+z)^2}{32\pi G k_B m_a \nu_0^2 H(z)}. \quad (5.29)$$

CoLoRe generates mock intensity mapping observations by modelling  $\delta_{\text{em}}$  as a biased tracer of  $\delta_M$  using one of the bias relations described in Section 5.2.4.1. The code estimates the mean brightness temperature in each Cartesian voxel in terms of the value of the matter overdensity, and a redshift-dependent mean temperature and bias. In order to interpolate from the Cartesian grid on to temperature maps, while accounting for the effects of redshift space distortions, each cell is first sub-divided into smaller sub-cells. The Cartesian coordinates of each sub-cell are translated into angular coordinates  $(\theta, \varphi)$  and emission frequency  $\nu = \nu_0/(1+z+v_r)$ , where  $z$  is the redshift corresponding to the sub-cell's radial comoving distance, and  $v_r$  is the value of the radial comoving peculiar velocity field. Each sub-cell is then assigned to a given pixel and frequency band based on  $(\theta, \varphi, \nu)$ . The final intensity maps are then generated by averaging over the temperature of all sub-cells thus assigned to each pixel.

The implementation used by CoLoRe is very similar to that used by CRIME (Alonso et al., 2014). CoLoRe does not emulate any other secondary effects, such as second-order gravitational lensing (Pourtsidou and Metcalf, 2015; Schaen et al., 2018) or self-absorption. If desired, gravitational lensing could be simulated

by perturbing the angular positions of the Cartesian sub-cells above with the deflection field described in Section 5.2.4.3. Fig. 5.6 shows a slice through a set of intensity maps corresponding to the 21cm line simulated using the model described in Alonso et al. (2014) in the range  $\nu \in [1015, 646]$  MHz ( $0.4 < z < 1.2$ ).

### 5.2.5 Lognormal predictions

Although, as discussed in Section 5.2.3.1, the log-normal transformation is not able to recover the right properties of the non-linear density fluctuations at all orders, one of its key advantages is the possibility to produce exact analytical predictions for the two-point correlators of a lognormal field. This makes it possible to compare the output of a CoLoRe simulation against precise predictions, making it straightforward to, for example, quantify the impact of different observational systematic effects added in post-processing on the main observables of a LSS experiment.

The real-space two-point correlation function of a lognormal field is related to that of its parent Gaussian field via:

$$1 + \xi_{\text{LN}}(r) = \exp [\xi_{\text{G}}(r)] , \quad (5.30)$$

where  $\xi_{\text{LN}}(r) \equiv \langle \delta_{\text{LN}}(\mathbf{x}) \delta_{\text{LN}}(\mathbf{x} + \mathbf{r}) \rangle$  and  $\xi_{\text{G}}(r) \equiv \langle \delta_{\text{G}}(\mathbf{x}) \delta_{\text{G}}(\mathbf{x} + \mathbf{r}) \rangle$ . Its cross-correlation with the Gaussian field itself is simply

$$\langle \delta_{\text{G}}(\mathbf{x}) \delta_{\text{LN}}(\mathbf{x} + \mathbf{r}) \rangle = \xi_{\text{G}}(r). \quad (5.31)$$

Computing the power spectrum of the lognormal field  $P_{\text{LN}}(k)$  from that of the parent Gaussian field  $P_{\text{G}}(k)$  is thus a simple three-step process:

1. Compute the Gaussian correlation function from  $P_{\text{G}}(k)$  via

$$\xi_{\text{G}}(r) = \frac{1}{2\pi^2} \int_0^\infty dk k^2 P_{\text{G}}(k) \frac{\sin(kr)}{kr}. \quad (5.32)$$

2. Compute  $\xi_{\text{LN}}$  from  $\xi_{\text{G}}$  through Eq. (5.30).

3. Compute the  $P_{\text{LN}}(k)$  from  $\xi_{\text{LN}}(r)$  via

$$P_{\text{LN}}(k) = 4\pi \int_0^\infty dr r^2 \xi_{\text{LN}}(r) \frac{\sin(kr)}{kr}. \quad (5.33)$$

CoLoRe automatically produces and outputs predictions for the power spectrum and correlation function of any lognormal biased tracers simulated. These



predictions are valid as long as the simulation was run using the lognormal structure formation model, and the exponential bias model (see Section 5.2.4.1), which retains the lognormal nature of the biased fields.

It is worth noting that, in order to produce truly accurate predictions, it is necessary to account for all sources of smoothing produced by the different operations carried out by CoLoRe:

- The chosen Gaussian smoothing scale.
- The finite resolution of the Cartesian grid.
- The effects of interpolating from the Cartesian grid onto beam-related quantities (line-of-sight integrals, skewers etc.).

The latter two effects (finite grid resolution and interpolation), have an associated Fourier-space window function given by

$$W_n(\mathbf{k}) = \left[ \left( 2 \frac{\sin(2k_x \Delta x)}{k_x \Delta x} \right) \left( 2 \frac{\sin(2k_y \Delta x)}{k_y \Delta x} \right) \left( 2 \frac{\sin(2k_z \Delta x)}{k_z \Delta x} \right) \right]^n, \quad (5.34)$$

where  $\Delta x$  is the grid spacing, and  $n = 1$  and  $2$  for nearest-neighbour and trilinear interpolation, respectively. For  $k \ll 1/\Delta x$ , we can approximate  $W_n$  as a Gaussian filter via

$$W_n(\mathbf{k}) \simeq \left[ 1 - \frac{(k_x^2 + k_y^2 + k_z^2)(\Delta x)^2}{24} \right]^n \simeq e^{-(kR_G)^2/2}. \quad (5.35)$$

where  $R_G^2 \equiv n(\Delta x)^2/12$ . Thus, the effective smoothing scale associated to the appropriate number of interpolation operations can be simply added in quadrature to the chosen Gaussian smoothing scale to produce theoretical predictions.

## 5.3 Results

### 5.3.1 Validation

CoLoRe has been the basis of other analyses, and some of its functionality was validated in previous work. The ability to produce large-scale intensity maps for the 21cm neutral hydrogen line was presented, used, and validated in Alonso et al. (2014) and Villaescusa-Navarro et al. (2017), and Witzemann et al. (2019) used the code to explore cross-correlations with galaxy clustering data. Farr et al. (2020b) used CoLoRe's line-of-sight skewers to produce mock observations of the Lyman- $\alpha$  forest; these mocks were extensively used to validate the final Lyman- $\alpha$  Baryon Acoustic Oscillations (BAO) analysis of the eBOSS collaboration (du Mas

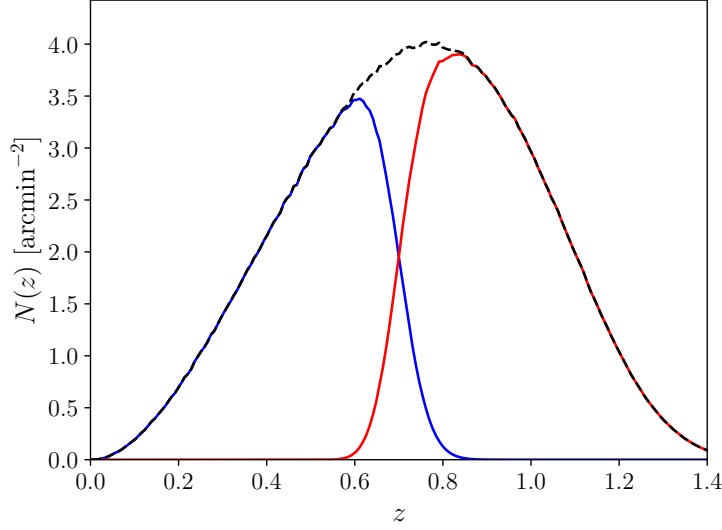


Figure 5.7: Redshift distribution of the two tomographic bins used in the analysis of the validation simulations (red and blue lines), as well as the overall redshift distribution (black dashed line). The bins are defined by a cut in photometric redshift space at  $z_{\text{photo}} = 0.7$ , where we assigned each source a random photometric redshift error with standard deviation  $\sigma_z = 0.03(1 + z)$ .

des Bourboux et al., 2020). Our discussion here therefore focuses on presenting and validating CoLoRe as a tool to produce fast simulations for wide galaxy surveys targeting galaxy clustering (both spectroscopic and photometric), weak lensing shear, and their cross-correlation with maps of the lensing convergence, and the ISW effect.

To do so, we have run a set of 100 large CoLoRe realisations containing these observables, and compared the relevant two-point correlations from different pairs of tracers in the simulation with the corresponding theoretical predictions.

### 5.3.1.1 Simulations

We generate 100 realisations covering the volume up to redshift  $z = 1.4$ , corresponding to a box size  $L_{\text{box}} = 5843 \text{ Mpc}/h$ . We use a grid of size  $N_{\text{grid}} = 2048$ , with cell size  $\Delta x = 2.85 \text{ Mpc}/h$ . Each realisation is populated with a single galaxy sample with the redshift distribution shown in Fig. 5.7, and a total number density  $\bar{n}_g = 2.9 \text{ arcmin}^{-2}$ . We used a lognormal structure formation model, and an exponential bias model with bias  $b(z) = 1 + 0.65z + 0.03z^3$ , compatible with a blue galaxy sample (Gabasch et al., 2006). We also used CoLoRe to generate maps of the lensing convergence and the ISW effect at  $z = 1$ . We used cosmological parameters  $\Omega_m = 0.3$ ,  $\Omega_b = 0.05$ ,  $h = 0.7$ ,  $n_s = 0.96$ ,  $\sigma_8 = 0.8$ . The Gaussian overdensity field was smoothed with a Gaussian kernel with width  $R_G = 2 \text{ Mpc}/h$ .

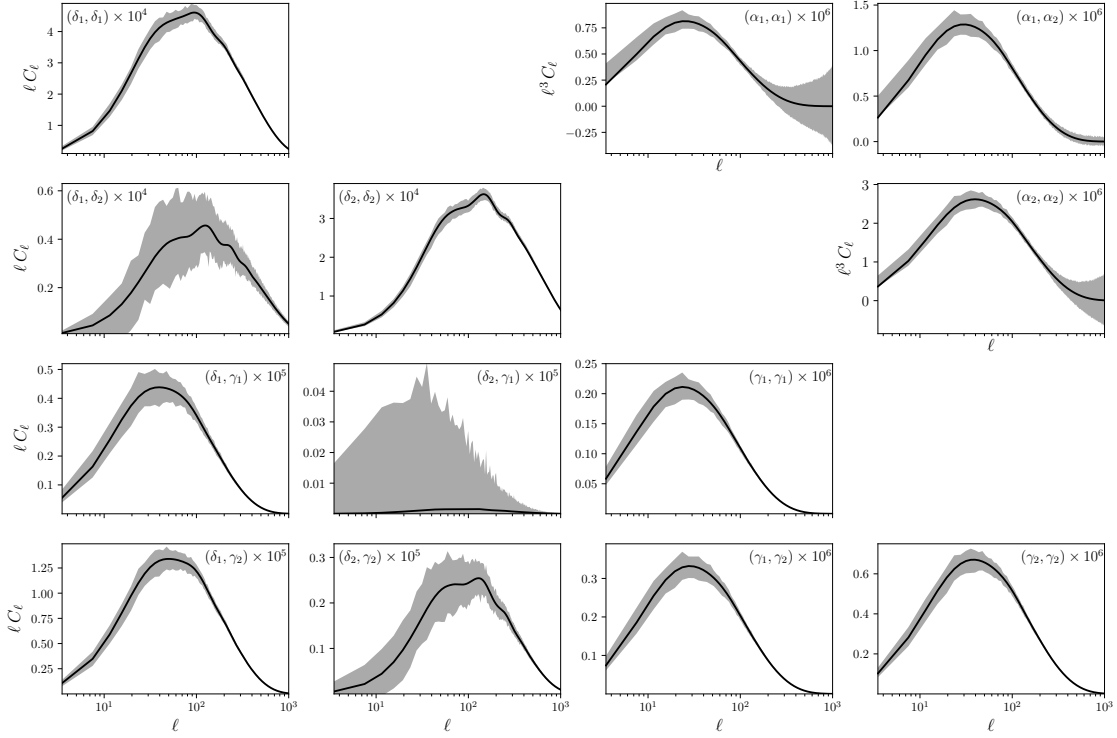


Figure 5.8: Galaxy clustering, cosmic shear and lensing displacement power spectra. The gray bands show the 68% scatter from the 100 validation realizations, and the theoretical predictions, described in Section 5.2.5, are shown as black solid lines. The lower-left corner shows all auto- and cross-correlations between the two galaxy clustering and cosmic shear bins. The upper-right corner shows the auto- and cross-correlations between the lensing displacement vectors in both redshift bins.

In order to simulate the effects of redshift uncertainties in photometric redshift surveys, we assigned a random Gaussian error to each galaxy redshift with standard deviation  $\sigma_z = 0.03(1+z)$ . We then divided all galaxies into two redshift bins, corresponding to sources above and below redshift  $z_{\text{thr}} = 0.7$ . The redshift distributions of the resulting bins are shown in Fig. 5.7. For each redshift bin, we created maps of the projected galaxy overdensity, as well as the shear and lensing displacement vectors. We did not make use of the fast lensing scheme described in Section 5.2.4.4 for these simulations. The latter quantities are provided by CoLoRe at each source. Finally, we computed all auto- and cross-power spectra between these maps, as well as the corresponding correlations with the lensing convergence and ISW maps at  $z = 1$  produced by CoLoRe. In order to optimally weight the shear and displacement fields by the local number of sources when computing these power spectra (and in order to account for their spin-2 and spin-1 nature), we make use of NaMaster (Alonso et al., 2019).



## CHAPTER 5. GENERATING SYNTHETIC DATASETS FOR COSMOLOGICAL PROBES WITH COLORE

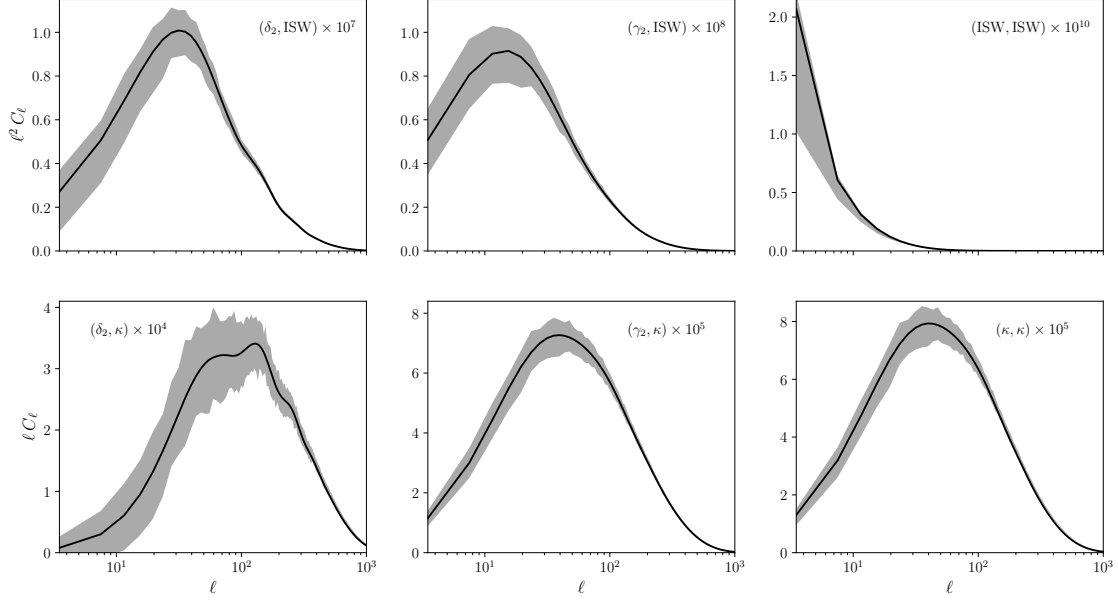


Figure 5.9: Cross-correlations between ISW (top row) and convergence maps (bottom row) at  $z = 1$  and the two high-redshift clustering and shear samples in the validation simulations. The gray bands show the 68% scatter from the 100 validation realizations, and the theoretical predictions, described in Section 5.2.5, are shown as black solid lines.

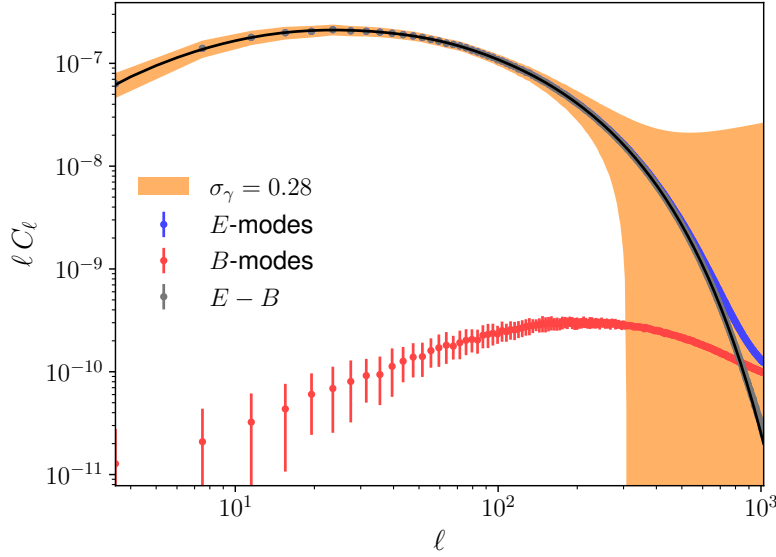


Figure 5.10: Shear power spectra for the first redshift bin in the validation simulations. The blue and red points show the results (mean and standard deviation of the 100 simulations) with no correction for the effects of source clustering. The gray points show the result of applying this correction by simply subtracting the  $B$ -mode power spectrum from the  $E$ -mode one. The black line shows the theoretical prediction for the latter. The orange band shows the  $1\sigma$  uncertainties one would find in the presence of realistic shape noise, which would make the source clustering effect undetectable in practice.

For the 3D clustering validation we used 10 out of the 100 realisations. Since the aim in this case is simulating a spectroscopic survey, we do not add photometric redshift uncertainties. We used the `corrfunc` (Sinha and Garrison, 2020) package to obtain measurements of the monopole and the quadrupole correlation functions for two different redshift bins (0.5, 0.7) and (0.7, 0.9); probing separations between  $r = 0.1 \text{ Mpc}/h$  and  $r = 200 \text{ Mpc}/h$  in 41 linearly spaced bins.

### 5.3.1.2 Validation of 2D observables

Before presenting the results from this validation exercise, it is worth clarifying a technical point about the theoretical predictions used to compare with the two-point functions estimated from the simulation. As described in Section 5.2.5, besides the effects of the lognormal transformation, one must account for the additional smoothing associated with the finite grid and the different interpolation operations. We do so by adding an extra smoothing scale in quadrature to the Gaussian smoothing scale used in the simulation, with the form

$$\Delta R_G^2 = n_{\text{eff}} \frac{(\Delta x)^2}{12}, \quad (5.36)$$

where the prefactor  $n_{\text{eff}}$  depends on the finite-resolution effects that must be taken into account. For instance, a single nearest-neighbour interpolation would correspond to  $n_{\text{eff}} \simeq 1$ , while linear interpolation would have  $n_{\text{eff}} \simeq 2$ . Thus, since galaxies are assigned to their nearest grid cell, and since two linear interpolations are needed to assign lensing properties to sources, the galaxy clustering and cosmic shear tracers produced by CoLoRe should take additional smoothing with  $n_{\text{eff}} \simeq 1$  and 4 respectively (not taking into account the additional interpolations involved in the fast lensing scheme). Since this additional Gaussian smoothing is not exact, we must adapt these prefactors slightly, in order to improve the agreement between theory predictions and simulations on small scales (high  $\ell$ s). For instance, we find that the galaxy-galaxy, galaxy-matter, and matter-matter power spectra must be smoothed by additional factors  $n_{\text{eff}}^{gg} = 0.9$ ,  $n_{\text{eff}}^{gm} = 4$ , and  $n_{\text{eff}}^{mm} = 3.8$  to recover the galaxy clustering and cosmic shear power spectra from the CoLoRe simulations. Therefore, care must be exercised when interpreting the results of CoLoRe simulations, particularly on physical scales smaller than, or comparable with, the grid resolution. The final users are encouraged to tune these  $n_{\text{eff}}$  factors to their particular analysis, depending on their use case and required level of accuracy.

The gray bands in Fig. 5.8 show the  $1\sigma$  scatter of the power spectra estimated from the 100 validation simulations, together with the corresponding theoretical

prediction in solid black. All auto- and cross-correlations between the projected galaxy overdensity and the cosmic shear field in the two redshift bins described above are shown in the lower-left panels in the figure. The upper-right panels show the three auto- and cross-correlations between the  $E$ -mode fields corresponding to the lensing displacement vectors in both redshift bins. Similar results are shown in Fig. 5.9 for correlations involving the lensing convergence and ISW maps, and the projected overdensity and cosmic shear fields in the second. A good agreement between theory and simulations, at the  $1\sigma$  level, is found in all cases.

One further complication must be noted. Cosmic shear is measured at the positions of clustered sources, and this clustering is correlated with the cosmic shear signal itself. This induces an additional contribution to the observed cosmic shear statistics which gives rise to shear  $B$ -modes, in addition to modifying the  $E$ -mode power spectrum. This effect is well known, and should be unobservable in most cases (Schneider et al., 2002). However, in the absence of shape noise (i.e. since we have access to the true weak lensing shear at each source), the effect can be measured in the catalogs produced by CoLoRe. This is illustrated in Fig. 5.10, which shows the  $E$  and  $B$ -mode power spectra for the first redshift bin in the validation simulations without any correction for source clustering (blue and red respectively) in comparison with the theory prediction without source clustering (black line) and the expected uncertainties in the presence of realistic shape noise (orange band). We find that, in practice, the effects of source clustering in the  $E$ -mode power spectrum can be corrected by simply subtracting the  $B$ -mode power spectrum from the  $E$ -mode power spectrum (gray points in the figure).

### 5.3.1.3 Validation of 3D clustering

We validated the 3D clustering of simulated galaxies both in real and in redshift space. The top panel of Fig. 5.11 shows the real-space monopole for two spectroscopic redshift bins: (0.5-0.7) and (0.7-0.9). The measurement of the correlation function was done separately in 48 different healpix pixels defining the CoLoRe beams, and the error bands (estimated from the scatter of these measurements) show the uncertainty for a single realization. Solid lines show the prediction derived in Section 5.2.5 with the linear bias left as a free parameter. The best-fit value of bias (using separations larger than 10 Mpc/h) agrees with the input value at the 2% level.

The bottom panel in Fig. 5.11 shows the redshift-space monopole and quadrupole for the same redshift bins and linear bias. We do not have a prediction for the clustering of galaxies in redshift space that is valid on all scales, but we use a modified

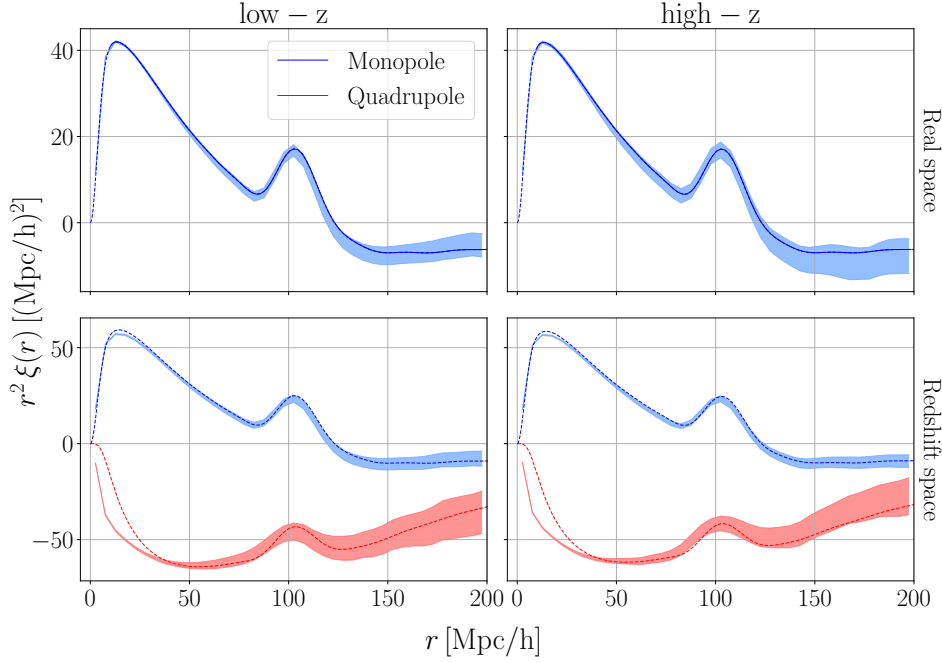


Figure 5.11: Measurements of the correlation function from the stack of 10 realizations used to validate the 3D clustering. The low- $z$  samples take redshifts from 0.5 to 0.7, while high- $z$  samples take redshifts from 0.7 to 0.9. The lines show the model, solid in the regions where the bias was fitted, and the shaded bands show the error for a single realization. *Top*: Measurements of the monopole in real space. *Bottom*: Measurements of the monopole and quadrupole in redshift space.

version of the lognormal model that includes linear Redshift Space Distortions (RSD) (Kaiser, 1987):

$$\delta_{\text{LN}}^s(\mathbf{k}) = \delta_{\text{LN}}(\mathbf{k}) + f\mu^2\delta_G(\mathbf{k}) \quad (5.37)$$

where the Gaussian term stands due to the fact that velocities comes directly from the gravitational potential (Eq. (5.25)). The redshift-space power spectrum is then:

$$P_{\text{LN}}^s(k, \mu) = P_{\text{LN}}(k) + f^2\mu^4P_G(k) + 2bf\mu^2P_G(k) \quad (5.38)$$

where we used Eq. (5.31) when computing the last term. Similarly to the 2D clustering, we added a smoothing associated with the finite grid with  $n_{\text{eff}} = 1$ . We also added an extra smoothing to correct for the binning of the correlation function measurement.

Differences on small scales between the measured and the predicted quadrupoles are due to inaccuracies in our RSD modelling, in particular higher-order terms ignored in Eq. (5.37). We discuss the role of these higher-order terms in Section 5.5, and show that they are indeed the cause of this disagreement.

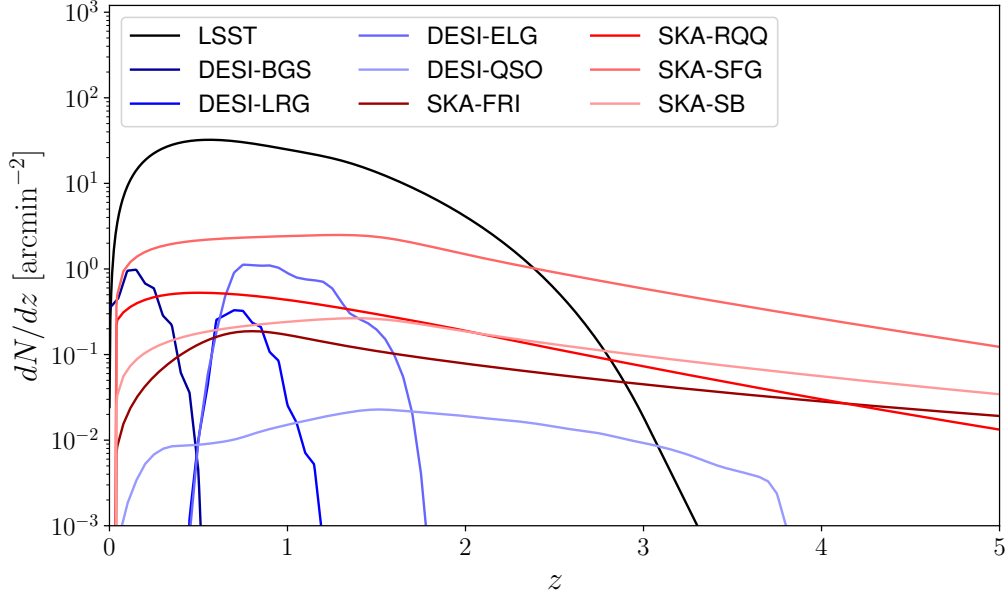


Figure 5.12: Redshift distribution of the different tracers simulated in our two flagship simulations. These include the LSST gold sample (black), the 4 DESI samples (blue) and 4 different radio continuum samples observable by SKA (red).

### 5.3.2 Performance at scale

In the context of existing and next-generation cosmological experiments, CoLoRe should be able to generate mock observations covering large volumes ( $z \lesssim 3$ ), with reasonable resolution ( $\Delta x = O(1)$  Mpc) for a wide range of observables (galaxy positions, shear, CMB lensing<sup>5</sup>, intensity maps, etc.). This Section quantifies the feasibility of these simulations.

#### Large-volume multi-tracer simulations

As an example of the type of mocks needed for Stage-IV surveys, we have used CoLoRe to generate two simulations containing cosmological probes for three major experiments:

- **DESI.** We simulate the Bright Galaxy Survey (BGS), Large Red Galaxies (LRG), Emission Line Galaxies (ELG), and quasi-stellar object (quasar) targets (including Ly- $\alpha$  skewers for the quasar sample) using the nominal  $N(z)$ , and bias functions from the DESI white paper (DESI Collaboration et al., 2016a). The resulting catalog contains  $\sim 3.4 \times 10^7$  BGSs,  $\sim 1.5 \times 10^7$  LRGs,  $\sim 10^8$  ELGs, and  $\sim 6.5 \times 10^6$  quasars over the full celestial sphere.

<sup>5</sup>Note that CoLoRe can provide the contribution to the CMB lensing convergence up to the highest redshift covered by the simulation box. Contributions from higher redshifts can then be added as a correlated Gaussian random field.

- **LSST.** We simulate a sample similar to the LSST "Gold" sample ( $i \lesssim 25.3$ ,  $\bar{n} \sim 40$  galaxies/arcmin<sup>2</sup>), resulting in  $\sim 6$  billion sources across the full sky. We follow the redshift distribution of the DESC Science Requirements Document (The LSST Dark Energy Science Collaboration et al., 2018), and assume a linear bias with redshift dependence  $b(z) = 0.95/D(z)$  (The LSST Dark Energy Science Collaboration et al., 2018; Nicola et al., 2020). The lensing shear, convergence and displacement is calculated for all sources using the fast lensing scheme described in Section 5.2.4.4.
- **SKA.** We simulate a radio continuum catalog comprised of 4 different radio galaxy samples: FRI radio-loud AGNs, radio-quiet AGNs (RQQs), normal star-forming galaxies (SFGs), and starbursts (SBs). For this we follow the models for the redshift distributions and linear bias described in (Wilman et al., 2008). The resulting samples contain  $\sim 4.5 \times 10^7$  FRIs,  $\sim 1.3 \times 10^8$  RQQs,  $\sim 8 \times 10^8$  SFGs, and  $\sim 8 \times 10^7$  SBs over the full sky. In addition to this, we generate simulated HI intensity mapping observations for 490 frequency bands covering the range  $473 \text{ MHz} < \nu < 947 \text{ MHz}$  (corresponding to redshifts  $0.5 < z < 2$ ). The maps were generated with an angular resolution  $N_{\text{side}} = 256$ , corresponding to pixels about 4 times smaller than the SKA primary beam in single-dish mode at the highest frequency ( $\theta_{\text{FWHM}} \sim 1.2^\circ$ ).

In addition to these, to showcase the ability of CoLoRe to generate CMB lensing observations, we generate a convergence map at resolution  $N_{\text{side}} = 1024$  at  $z = 3$ , caused by the same matter density field that serves as seed for the tracers listed above. The redshift distributions of the different galaxy samples simulated are shown in Fig. 5.12. For illustrative purposes, Fig. 5.13 shows the 3-dimensional representation of some of the quantities simulated by CoLoRe in one of the beams used by the code.

The simulations were generated with a  $\Lambda$ CDM model compatible with the best-fit *Planck* cosmological parameters (Planck Collaboration et al., 2020). The boxes span the redshift range ( $0 < z < 3$ ), with  $N_{\text{grid}} = 4096$ , resulting in a spatial resolution of  $\approx 2 \text{ Mpc}/h$ .

### Run time, memory usage, and fast lensing

Both simulations were initialised with the same random seed, but using different structure formation models, lognormal (LN) and first-order LPT (1LPT) respectively. Both simulations were run at NERSC <sup>6</sup>. The LN simulation was generated

<sup>6</sup><https://nersc.gov>

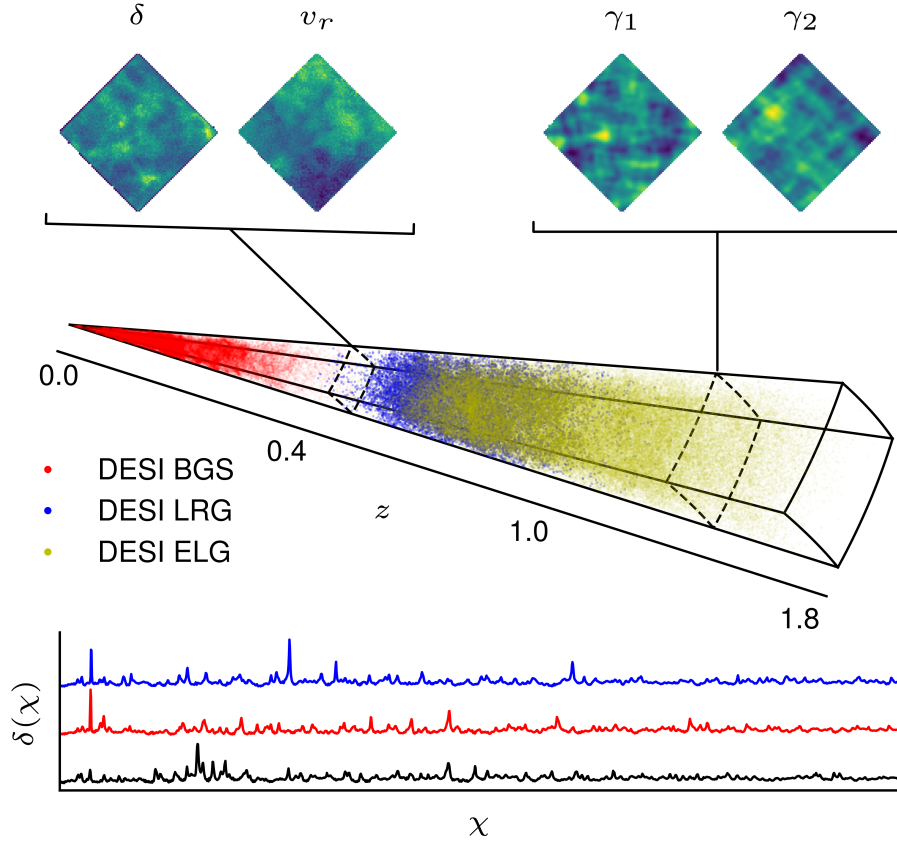


Figure 5.13: Visual description of the multi-tracer products that can be simulated with CoLoRe. The upper plot shows one of the beams used internally by CoLoRe for domain decomposition, with the redshift and angular coordinates of 3 different DESI samples, and maps of the density, radial velocity, and lensing shear constructed from sources at two different redshifts. The lower plot shows the density skewers calculated for three arbitrary DESI quasars contained in the same beam, as a function of comoving distance.



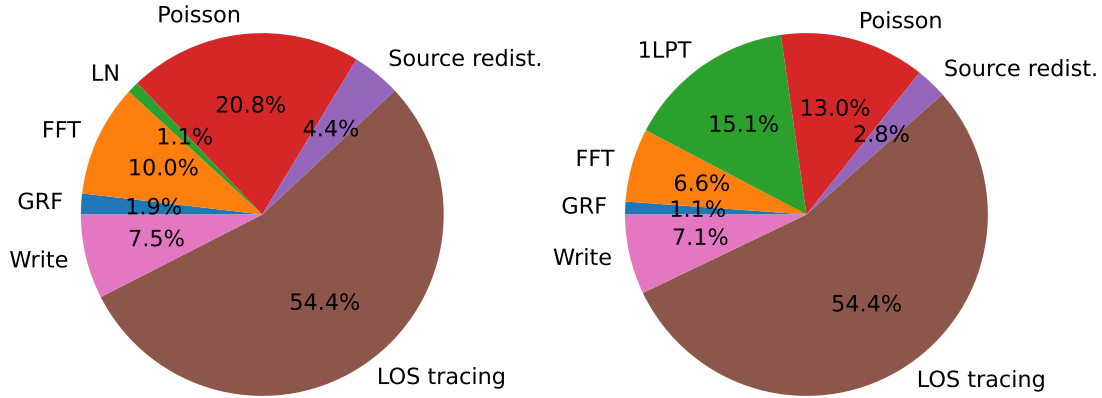


Figure 5.14: Fraction of the total run time taken up by different stages of a typical CoLoRe simulation. The stages shown are, the generation of the initial Gaussian random fields in Fourier space ("GRF"), their transformation to real space ("FFT"), the structure formation model leading to a positive-definite matter overdensity ("LN" or "1LPT" for the lognormal and first-order LPT simulations), the generation of source catalogs via Poisson sampling ("Poisson"), the redistribution of these sources across different nodes before any line-of-sight calculations ("Source redistrib."), the calculation of all relevant line-of-sight quantities ("line of sight tracing"), and the output of all final products to disk ("Write"). Note that the line of sight tracing stage is more sensitive to inter-node communication than other stages, and thus it takes up the same fraction of the total compute time in both simulations in spite of the additional time taken by the 1LPT stage, given the larger number of MPI nodes needed for that simulation.

using 40 MPI tasks, distributed across 20 Cori-haswell nodes, using 16 OMP threads per task. This simulation ran in 1.15 hours, using approximately 730 CPU-hours. The 1LPT simulation required a larger number of nodes, given the additional memory needed to allocate the three more Cartesian grids mentioned in Section 5.2.3.2. This simulation was run using 72 MPI tasks distributed across 36 Cori-haswell nodes, using 16 OMP threads per task. The simulation ran in 0.93 hours using a total of 1 075 CPU-hours. A suite of 1 000 such simulations could therefore be run using  $\sim 1$  million CPU-hours.

Table 5.1 lists the memory and disk requirements associated with each of these tracers<sup>7</sup>. Although the most memory-demanding task is the generation of the 3D density and Newtonian potential fields, or the Lagrangian displacement components if using LPT, the final data products also lead to a non-negligible

<sup>7</sup>Note that, naively, we have simulated all galaxy tracers as disjoint samples when, in reality, there would be significant overlap between them (e.g. between SKA SFGs and the LSST sample). A more realistic setting should therefore account for these overlaps when generating the different galaxy catalogs.



## CHAPTER 5. GENERATING SYNTHETIC DATASETS FOR COSMOLOGICAL PROBES WITH COLORE

Tracer	$\Delta z_{\text{RSD}}$	$\alpha$	$\gamma$	$\kappa$	$\delta(z), v_r(z)$	Memory (GB)	Disk (GB)
LSST	✓	✓	✓	✓	✗	308	155
DESI-BGS	✓	✗	✗	✗	✗	1.7	0.6
DESI-LRG	✓	✗	✗	✗	✗	0.8	0.3
DESI-ELG	✓	✗	✗	✗	✗	5.2	1.9
DESI-QSO	✓	✗	✗	✗	✓	99	99
SKA-FRI	✓	✗	✗	✗	✗	2.3	0.8
SKA-RQQ	✓	✗	✗	✗	✗	6.8	2.5
SKA-SFG	✓	✗	✗	✗	✗	39	15
SKA-SB	✓	✗	✗	✗	✗	4.2	1.5
SKA-21cm	✓	N.A.	N.A.	N.A.	N.A.	2.8	1.4
Convergence map	N.A.	N.A.	N.A.	✓	N.A.	0.1	0.05
$\delta, \phi_N$ grids	N.A.	N.A.	N.A.	N.A.	N.A.	768	N.A.
$\Psi_{\text{1LPT}}$ grids	N.A.	N.A.	N.A.	N.A.	N.A.	1229	N.A.

Table 5.1: Simulation products generated by CoLoRe. The first 11 rows show the different tracers generated for the large-volume simulations described in the text. For each tracer we show the physical quantities simulated (RSD, lensing displacements, shear, convergence, and density/velocity skewers), as well as the memory and disk space taken. The last two rows display the memory requirements for the different Cartesian grids stored for lognormal and 1LPT simulations. Although the memory requirements are dominated by these cartesian grids (particularly for LPT simulations), the simulated tracers can take up a non-negligible fraction of the available memory. This is patent for the LSST sample, given its size, and the need to store lensing information, and for the DESI quasar sample, since we save a full density/velocity skewer for each source.

memory requirement, particularly in the case of high-density samples such as LSST, or for a large number of density/velocity skewers ( $\sim 27\%$  and  $\sim 4\%$  respectively in the case of the lognormal simulation).

Fig. 5.14 shows the time taken by the different stages in both runs. In both cases the slowest stage is the collection of line-of-sight information, such as the density and velocity skewers stored for all DESI quasars, or the lensing information associated with the LSST sources. Given the large number of sources in the LSST sample, this stage would completely dominate the run time if we had not used the “fast lensing” method described in Section 5.2.4.4.

To quantify this, we ran two additional identical simulations with a total of 5.9 billion galaxies, emulating a blue galaxy population of LSST 10-year depth. Both simulations were generated using 32 MPI tasks distributed across 16 Cori Haswell nodes at NERSC. Each MPI task had 16 OMP threads. The simulation using the fast lensing scheme took  $\sim 35$  minutes to run (a total of 173 CPU-hours), while the

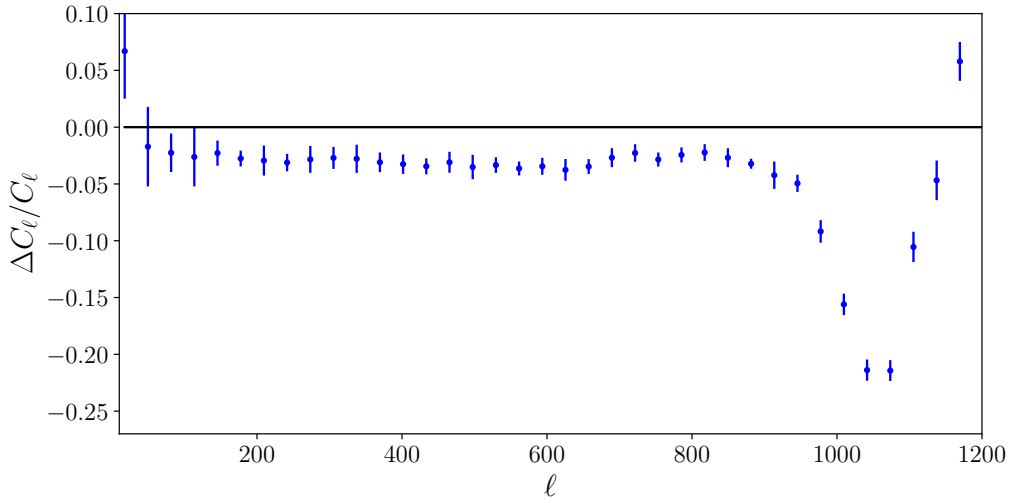


Figure 5.15: Relative difference between the shear power spectrum of a CoLoRe simulation run using the "fast lensing" scheme, and a simulation run without this approximation. The effects caused by the various interpolations carried out as part of the fast scheme should be carefully modelled if an accurate theoretical description of the CoLoRe output is required.

one integrating along each galaxy's line of sight took  $\sim 36.3$  hours (i.e.  $\sim 18,600$  CPU-hours), more than 60 times longer. The additional time was completely taken by the line of sight calculations. The fast-lensing scheme thus allows for a factor  $\sim 60$  speed-up and, in the current implementation of CoLoRe, is absolutely necessary for simulations with billions of sources.

The price to pay for this speed-up is the additional complexity associated with the interpolation operations involved in the fast scheme (interpolation from the cartesian grid to the fixed pixel lines-of-sight, and from those to the galaxy positions). Fig. 5.15 shows the relative difference between the cosmic shear power spectrum computed from a simulation run with the fast lensing scheme, and one run without this approximation. The effects of the fast lensing approximation are a decrease in the final lensing amplitude, and a gradual loss of power at higher multipoles, on scales comparable with the size of the adaptive pixels used in this approximation. These effects would need to be accurately characterized if the application of the CoLoRe realizations requires an accurate theoretical prediction of two-point statistics involving weak lensing observables. The accuracy/speed-up trade-off can be controlled by the user by changing the number of shells in which to compute the lensing information, as well as the size of the adaptive pixels in relation with the Cartesian cell size.

## 5.4 Conclusions

In this Chapter we have introduced CoLoRe, a public code to efficiently generate synthetic realisations of multiple cosmological surveys. We started in Section 2 by describing the overall structure of the code, and the different methods to simulate the density field. We have presented the available tracers in CoLoRe, and how to add new ones using its highly modular structure. We concluded this Section discussing the accurate predictions available when working with the lognormal model of structure formation.

In Section 3 we presented the validation results of some of the key summary statistics from the simulated maps, using a large set of CoLoRe boxes. We showed that the measured angular power spectra from simulated photometric surveys agree with the theoretical predictions up to  $\ell = 1000$ . This is true for galaxy correlations, galaxy-shear cross-correlations and several lensing statistics (shear, convergence, displacements). The 3D clustering in simulated spectroscopic galaxy surveys was also validated in the absence of redshift space distortions, where the lognormal model can be accurately predicted on all scales. As discussed in Section 5.5 we do not have a good model for the small-scales multipoles in redshift space, but the agreement is very good on large, linear scales.

We have discussed the performance of CoLoRe at scale by presenting two joint simulations of DESI, LSST and SKA. These large boxes cover the whole comoving volume out to  $z = 3$ , with a resolution of  $\approx 2\text{Mpc}/h$ , and include spectroscopic and photometric galaxies, lensing, intensity mapping and radio galaxies. The more realistic simulation, using Lagrangian Perturbation Theory (LPT), only used about 1 000 CPU-hours. Simulating hundreds of these boxes is entirely feasible, and can be used to characterise systematic effects in multi-experiment analyses, or estimate cross-survey covariances.

Finally, we discussed the differences between the two options used in CoLoRe to compute weak lensing variables from source galaxies. The fast lensing implementation provides a factor of  $\sim 60$  speed-up for an LSST-like sample, while maintaining an accuracy better than a 2-3% bias in the amplitude of the shear power spectrum on large scales ( $\ell < 1000$ ).

There are several features that could be added to CoLoRe without major changes in the code structure:

- **Better structure formation:** Currently CoLoRe can simulate the growth of structure using a lognormal model or LPT computed at first or second order. Future versions of the code could add new modules to use more

complex models of the growth of structure, such as COmoving Lagrangian Acceleration (COLA) (Tassev et al., 2013) as in Izard et al. (2017), where it was used to generate weak lensing maps and halo catalogues in the lightcone; or Fast Particle-Mesh (FastPM; Feng et al., 2016)) algorithms. The potential additional compute time and memory requirements associated with these methods should be weighed against the need for more accurate clustering statistics of a given application.

- **Non-linear RSD:** Minor modifications of the code could improve the level of realism of the RSD, by sourcing the velocities from the computed LPT fields. This change could have a significant impact on the 3D clustering of galaxies intermediate scales. However, CoLoRe does not currently simulate virialized objects, and therefore we are not able to properly capture non-linear peculiar velocities or Fingers of God. Random virial motions (or redshift errors) can be added in post-processing, by assigning random shifts to the galaxy redshifts, but a more realistic approach of non-linear RSD is beyond the scope of this work.
- **Halos:** The models currently used by CoLoRe to generate different tracer observations directly connect the latter with the underlying smooth density field. The complexity and fidelity of these simulations could be improved if this density field was endowed with a halo catalog. This could be done directly at the level of the linear density field using Press-Schechter-inspired methods (as in e.g. Santos et al. (2010)), or from the LPT displacement field using a modified friends-of-friends search (Manera et al., 2013), or peak-patch methods (Stein et al., 2019, 2020). A halo catalog would allow us to improve the fidelity of the resulting density field on small scales by including the expected density profile of these halos, and would make it possible to use the halo model to generate simulated observations of additional tracers (e.g. thermal or kinematic Sunyaev-Zel’dovich effects, halo-occupation distributions for better galaxy catalogs etc.).
- **Small areas and flat-skies:** The current version of CoLoRe simulates the whole universe to a given redshift, with the observer placed in the center of a large box. Users interested in simulating surveys with relatively small areas might prefer to place the observer in one side of rectangular box and simulate only a (literal) light-cone. For sufficiently small sky areas, these simulations could be made faster making use of the flat-sky approximation. Both of these features should be easy to implement in CoLoRe.

The addition of these features should not significantly impact the performance of CoLoRe reported here. By making our code public we want to encourage the different collaborations preparing the next generation of large cosmological surveys to use CoLoRe to efficiently generate realistic synthetic simulations of their datasets to be used in multi-survey analyses.

## 5.5 Appendix: Higher-order terms in the modelling of redshift-space distortions

The redshift-space galaxy overdensity,  $\delta_{\text{LN}}^s = \delta_{\text{LN}}(\mathbf{s})$ , is related to its real-space equivalent  $\delta_{\text{LN}}(\mathbf{x})$  and the normalized gradient of line-of sight velocities  $\eta = -\partial_z v_z / H(z)$  via:

$$1 + \delta_{\text{LN}}^s = \frac{1 + \delta_{\text{LN}}}{1 - \eta} \quad (5.39)$$

Assuming Gaussian RSDs are small, we can expand this in powers of  $\eta$ ,

$$\delta_{\text{LN}}^s = \delta_{\text{LN}} + \eta + \epsilon, \quad (5.40)$$

where  $\epsilon(\mathbf{x}) \equiv \delta_{\text{LN}}(\mathbf{x}) \eta(\mathbf{x})$ . This term would be order-2 assuming  $\delta_{\text{LN}}$  is small, but we do not make this approximation here. Ignoring this term one would recover Eq. (5.37), used for the theoretical predictions in Section 5.3.1.3. If we keep this extra term, however, the model for the redshift-space galaxy power spectrum will have new contributions with respect to the model described in Eq. (5.38):

$$\begin{aligned} P_{\text{LN}}^s(k, \mu) &= (\Delta k)^3 \langle |\delta_{\text{LN}}^s(\mathbf{k})|^2 \rangle \\ &= P_{\text{LN}}(k) + f^2 \mu^4 P_{\text{G}}(k) + 2bf\mu^2 P_{\text{G}}(k) \\ &\quad + (\Delta k)^3 [2\langle \delta_{\text{LN}}(\mathbf{k}) \epsilon(\mathbf{k}) \rangle + 2\langle \eta(\mathbf{k}) \epsilon(\mathbf{k}) \rangle + \langle \epsilon(\mathbf{k}) \epsilon(\mathbf{k}) \rangle] . \end{aligned} \quad (5.41)$$

Even though it is possible to compute analytical predictions for the correlation function including these new  $\epsilon$  terms (as convolutions of  $\delta_{\text{LN}}$  and  $\eta$  in Fourier space), we leave this for future work. Here we only quantify each of the terms by looking at cross-correlations of test galaxy samples from a custom CoLoRe simulation designed for this study. This simulation was similar to those described in Section 5.3.1.1 but with two spectroscopic galaxy samples: a first sample  $\delta_A$  with an extremely low clustering amplitude ( $b_A = 0.001$ ), and a second sample  $\delta_B$  with a redshift-independent bias of  $b_B = 2$ .

While the clustering of sample  $A$  in real space is negligible, its redshift-space power spectrum can be modeled as  $\langle \delta_A^s(\mathbf{k}) \delta_A^s(\mathbf{k}) \rangle \propto f^2 \mu^4 P_{\text{G}}(k)$  (up to a normalisation factor  $(\Delta k)^3$ ). This can be clearly seen in the (a) panel of Figure Fig. 5.16,

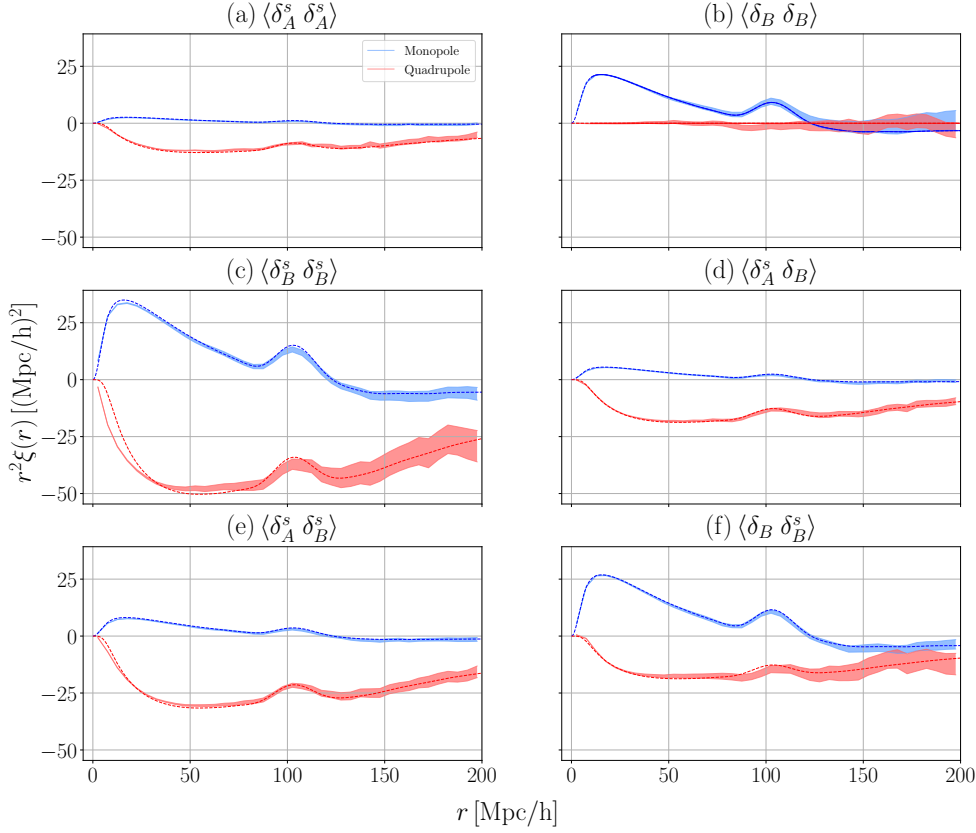


Figure 5.16: Measurements and predictions for the different cross-correlations discussed in Section 5.5 for a single simulation with two special tracers ( $b_A = 0.001$  and  $b_B = 2$ ). The predictions are from the simpler model discussed in Section 5.3.1.3 and are missing those terms involving  $\epsilon(\mathbf{x}) = \delta_{LN}(\mathbf{x})\eta(\mathbf{x})$ . This explains the disagreement seen on the small-scales quadrupole in panel (c), it also shows a small disagreement in panels (e) and (f).

and validates the simulation of RSDs in CoLoRe. The clustering of sample  $B$  in real (redshift) space is shown in the (b) ((c)) panels of the same figure, and are similar to Fig. Fig. 5.11 discussed in Section 5.3.1.3. While its real space clustering is well described by the lognormal model  $\langle \delta_B(\mathbf{k})\delta_B(\mathbf{k}) \rangle \propto P_{LN}(k)$ , the small-scales quadrupole of its redshift-space equivalent can not be described by the simple model of Eq. (5.38). We are missing the contributions from the terms  $\langle \delta_{LN} \epsilon \rangle$ ,  $\langle \eta \epsilon \rangle$  and  $\langle \epsilon \epsilon \rangle$  introduced in Eq. (5.41).

In panel (d) we show the cross-correlation of  $\delta_A^s$  and  $\delta_B$ , compared to its prediction  $\langle \delta_A^s(\mathbf{k})\delta_B(\mathbf{k}) \rangle \propto b_B f \mu^2 P_G(k)$ . Because there is no  $\epsilon$  term involved in this cross-correlation, the model describes the measurement very well on all scales. In panel (e) we cross-correlate  $\delta_A^s$  with  $\delta_B^s$  instead, and compare it to a theoretical

prediction that includes the two Kaiser terms ( $b_B f \mu^2 P_G(k) + f^2 \mu^4 P_G(k)$ ) but is missing an extra term  $\langle \eta \epsilon \rangle$ . The minor disagreement of the quadrupole on small scales allows us to estimate the magnitude and sign of the missing term. Finally, in (f) we show the cross-correlations of  $\delta_B$  and  $\delta_B^s$ , i.e., the cross-correlation of the same B sample in real and in redshift space. We plot its prediction from the simpler model from Section 5.3.1.3, including the terms  $P_{LN}(k) + b_B f \mu^2 P_G(k)$ . Again, following Eq. (5.41) this cross-correlation should include an extra term  $\langle \delta_{LN} \epsilon \rangle$  that explains the small disagreement of the quadrupole on small scales.

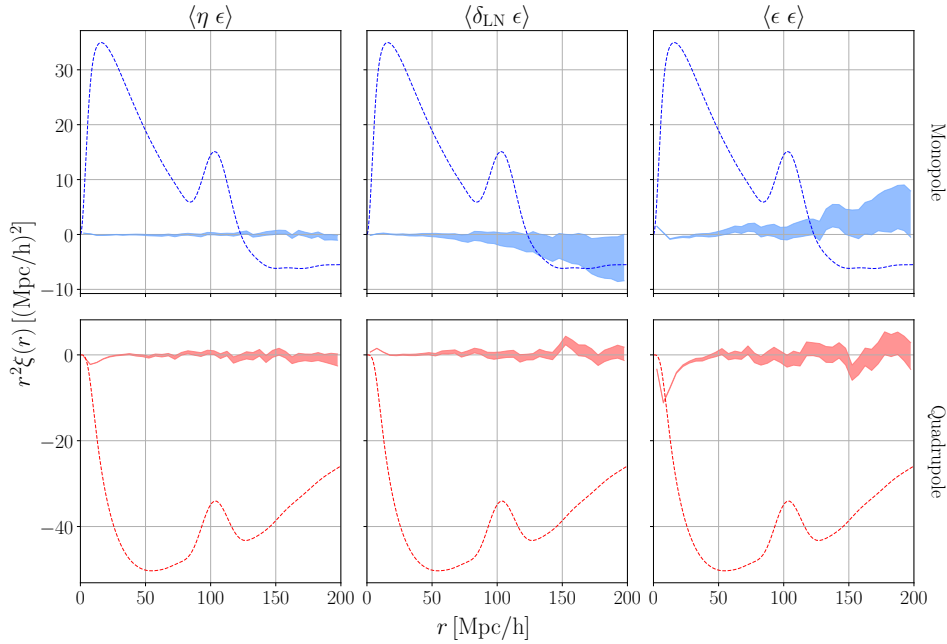


Figure 5.17: Contributions to the monopole (top panel/blue lines and bands) and the quadrupole (bottom panel/red lines and bands) from the terms in Eqs. Eq. (5.42)-Eq. (5.44) using the same simulation as in Figure Fig. 5.16. The shaded bands show the error in the measurement of each term from the scatter between 48 healpixels of  $N_{side} = 2$ . The dashed lines show the full model prediction as in Section 5.3.1.3, where the extra terms are not included. Particularly the term  $\langle \epsilon \epsilon \rangle$  (right panel) has an important impact on the small-scales quadrupole.

Following Eq. (5.39) it is clear that  $\epsilon = \delta_B^s - \delta_B - \delta_A^s$ . Therefore, by combining the different cross-correlations discussed above we can now isolate each of the new terms in Eq. (5.41):

$$\langle \delta_{LN} \epsilon \rangle = \langle \delta_B \delta_B^s \rangle - \langle \delta_B \delta_B \rangle - \langle \delta_B \delta_A^s \rangle \quad (5.42)$$

$$\langle \eta \epsilon \rangle = \langle \delta_A^s \delta_B^s \rangle - \langle \delta_A^s \delta_B \rangle - \langle \delta_A^s \delta_B \rangle \quad (5.43)$$

$$\langle \epsilon \epsilon \rangle = \langle \delta_B^s \delta_B^s \rangle + \langle \delta_B \delta_B \rangle + \langle \delta_A^s \delta_A^s \rangle + 2\langle \delta_A^s \delta_B \rangle - 2\langle \delta_A^s \delta_B^s \rangle - 2\langle \delta_B \delta_B^s \rangle. \quad (5.44)$$

The multipoles corresponding to these three combinations of correlations are shown in Figure Fig. 5.17, together with the model prediction from equation Eq. (5.38). One can see that the largest correction to the model should come from the  $\langle \epsilon \epsilon \rangle$  term, while the other two terms are small and partially cancel each other.





# THE LYMAN- $\alpha$ FOREST CATALOG FROM THE DARK ENERGY SPECTROSCOPIC INSTRUMENT EARLY DATA RELEASE

In the previous Chapter, we dealt with mock simulations, described in the context of cosmological analyses as an aid of the analysis and understanding of real data. In contrast, in the next Chapters we will focus on the Lyman- $\alpha$  forest analyses for the Dark Energy Spectroscopic Instrument (DESI) survey.

In this Chapter, we will present and validate the catalog of Lyman- $\alpha$  forest fluctuations for 3D analyses using the Early Data Release (EDR) from the DESI survey. This catalog contains information from 88 511 quasars collected from Survey Validation (SV) data and the first two months of the main survey (M2).

The text from this Chapter is extracted from my publication Ramírez-Pérez et al. (2024), published in the Monthly Notices of the Royal Astronomical Society in March 2024.

The structure of this Chapter will be as follows. In Section 6.1, we will provide more context to the Lyman- $\alpha$  forest analyses, fitting them into current cosmological research. In Section 6.2, we present the data used for this work, including spectra and quasar catalogs. In Section 6.3, we explain how we obtain the flux-transmission field from spectra through the estimation of the expected flux of quasars in the continuum fitting process. This includes masking some wavelengths and applying corrections to the flux calibration and the reported uncertainties by the pipeline. In Section 6.4, we discuss aspects that require further clarification, such as modified weights and wavelength grid choices. Finally, in Section 6.5, we provide a summary.

## 6.1 Introduction

Measurements of Lyman- $\alpha$  forest 1D correlations in a handful of high-resolution quasar spectra emerged as a powerful tool to study the large-scale distribution of matter (Croft et al., 1998; McDonald et al., 2000), opening a new field in the analysis of the high redshift universe and helping to constrain cosmological parameters.

Using data from the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al., 2013), the three-dimensional correlation function of absorption in the Lyman- $\alpha$  forest was measured for the first time in Slosar et al. (2011). Shortly after that, the first measurement of the Baryon Acoustic Oscillations (BAO) peak in the Lyman- $\alpha$  forest was presented (Busca et al., 2013; Kirkby et al., 2013; Slosar et al., 2013), using data from BOSS DR9 (Lee et al., 2013). These were followed by other BAO analyses using increasingly larger Lyman- $\alpha$  forest datasets from BOSS (Delubac et al., 2015; Bautista et al., 2017) and from the Extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al., 2016; de Sainte Agathe et al., 2019). The precision of these BAO measurements was significantly improved with the measurement of the cross-correlation of quasars and the Lyman- $\alpha$  forest (Font-Ribera et al., 2014; du Mas des Bourboux et al., 2017; Blomqvist et al., 2019), and the final Lyman- $\alpha$  BAO measurement combining BOSS and eBOSS was presented in du Mas des Bourboux et al. (2020).

DESI is currently undergoing a five-year campaign to obtain close to a million quasar spectra with  $z > 2$  (Chaussidon et al., 2023; DESI Collaboration et al., 2023b). This dataset will be four times larger than the state-of-the-art (the eBOSS DR16 quasar sample presented in Lyke et al. (2020)), and will enable sub-percent BAO measurements with the Lyman- $\alpha$  forest (Levi et al., 2013; DESI Collaboration et al., 2016a).

We present the first catalog of Lyman- $\alpha$  forest fluctuations in DESI, including data from the Early Data Release (EDR; DESI Collaboration et al., 2023a) and from the first two months of the main survey (M2). This dataset was used in Gordon et al. (2023) for the first measurement of 3D correlations in the Lyman- $\alpha$  forest from DESI. It was also used in Herrera et al. (2023) in a comparison with synthetic datasets.

The methodology used here is similar to the one developed for eBOSS analyses, especially the most recent analysis by du Mas des Bourboux et al. (2020). This served as the basis for developing the data analysis pipeline of the DESI Lyman- $\alpha$  forest working group. In this work, we provide a detailed description of our new pipeline, focusing on the changes with respect to the one used in eBOSS analyses. Some of these changes are motivated by changes in the input data: for instance,

while SDSS spectra had pixels equispaced in the logarithm of the wavelength, DESI uses linearly spaced pixels. Other changes are motivated by studies that appeared after du Mas des Bourboux et al. (2020). For instance, following Ennesser et al. (2022) we now include in our analysis the spectra of Broad Absorption Line (BAL) quasars, with the most contaminated regions properly masked. Finally, we also revisit the weighting scheme used to compute correlations in the Lyman- $\alpha$  forest, resulting in an improvement of more than 20% in our precision.

All of the process followed here is executed using the publicly available code `PICCA`<sup>1</sup> and can be reproduced by the user using public DESI data. The `PICCA` package also includes modules for the computations of both auto- and cross-correlation with quasars, cosmological fits, and multiple useful tools for Lyman- $\alpha$  forest studies.

The catalog is aimed at studies of 3D correlations in the Lyman- $\alpha$  forest. A similar dataset is used by measurements of the 1D correlations with DESI data (Ravoux et al., 2023; Karaçaylı et al., 2024). Although analogous estimations of the unabsorbed quasar continuum are also needed in these studies, their methodology is somewhat different and these publications focus on studies of systematics that primarily affect the 1D correlations.

## 6.2 Data

Data used in this work comes from two different DESI datasets. On the one hand, we have the Early Data Release (EDR), which includes all Commissioning, Survey Validation (SV), and special survey data. On the other hand, we have EDR+M2, which includes all data from EDR as well as the first two months of the main survey.

Although this two samples are qualitatively similar, we are performing the analysis separately for each of them to achieve two purposes:

- **Describe the Lyman- $\alpha$  forest Value Added Catalog (VAC) in the context of EDR+M2 (DESI Collaboration et al., 2023a):** VACs for a wide variety of tracers are being released using DESI early data. The present work provides the Lyman- $\alpha$  forest fluctuations catalog.
- **Describe the Lyman- $\alpha$  forest catalog used in the context of early DESI data publications:** Multiple related publications are being published in this context within the Lyman- $\alpha$  working group. The objective of Gordon et

<sup>1</sup><https://github.com/igmhub/picca/>

al. (2023) is to obtain the first Lyman- $\alpha$  correlation measurements from DESI early data, testing the current pipeline and data quality, and compare its performance to previous eBOSS DR16 analyses. Herrera et al. (2023) provides details on the current status of the different procedures used to build mocks for Lyman- $\alpha$  forest analyses. Gontcho et al. (2023) characterizes the systematics caused by the DESI instrument on the 3D correlations of the Lyman- $\alpha$  forest. Bault et al. (2023) studies the impact of redshift errors on the 3D cross-correlation of quasars with the Lyman- $\alpha$  forest. P1D analyses are performed in two different papers: Ravoux et al. (2023) presents the 1 dimensional measurement using Fast Fourier Transform (FFT), while Karaçaylı et al. (2024) makes use of the Quadratic Maximum Likelihood Estimator (QMLE).

Finally, the current work provides the Lyman- $\alpha$  fluctuations catalog, as well as its validation. The decision to use EDR+M2 data for these analyses was motivated by the need for a larger volume of data than the Early Data Release plus the first two months of the main survey (EDR+M2) could offer, which leads to better measurements of the correlation function, constraining power and estimation of systematics.

EDR+M2 data, including Lyman- $\alpha$  forest fluctuations is available now, including this VAC describing Lyman- $\alpha$  fluctuations. However, EDR+M2 will not be released as a separate piece of data and M2 will be released alongside Data Release 1 (DR1) data.

### 6.2.1 DESI spectroscopic data

Data from the DESI spectroscopic pipeline comes divided in files where the information of multiple observations is coadded. Each coadded file contains per-spectra information for the three arms (B, Z, R), as well as various metadata, including fiber positions, instrument configuration during observation and atmospheric conditions. The spectroscopic data include fluxes, estimated inverse-variance and a mask identifying invalid pixels.

Each of the three arms of the spectrograph covers a different wavelength region:

- B arm: [3600, 5800] Å
- R arm: [5760, 7620] Å
- Z arm: [7520, 9824] Å

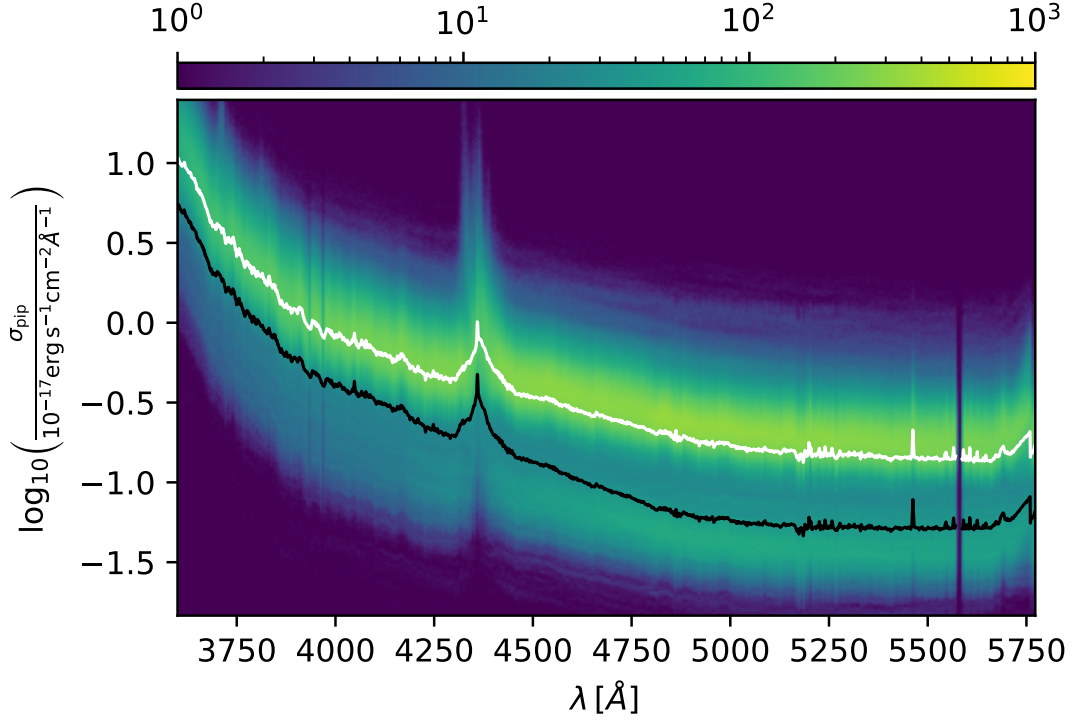


Figure 6.1: Distribution of wavelength and pipeline-reported error on the flux measurement for the larger EDR+M2 sample. This measurement is performed in the C III region of the spectra, as defined in Table 6.2. The solid lines mark the mean value of the error for a given wavelength, black for the EDR sample and white for the EDR+M2. Apart from the clear distinction in these two lines, a subtle division into two bands at larger wavelengths can also be observed. This is caused by the better signal-to-noise in the EDR sample, thanks to multiple re-observations of the same targets. We can also observe how the variance relatively increases in the two ends of the spectrograph and in the area affected by the collimator mirror reflectivity around 4400 Å.

with some overlap between each of the arms. The Lyman- $\alpha$  forest is predominantly observed in the blue arm (B arm) of the spectrograph, and only partially in the red arm (R arm). DESI follows a linear wavelength grid with 0.8 Å steps, PICCA is adapted to work with this resolution.

In Fig. 6.1, we show the distribution of wavelength and pipeline-reported error on the flux measurements for the EDR+M2 dataset, where  $\sigma_{\text{pip}}$  the variance reported by the DESI pipeline based on photon statistics and readout noise. The image highlights the main differences between EDR+M2 and M2 samples. EDR+M2 data corresponds to longer observations, with fewer quasars observed. M2 data, on the other hand, includes a large number of objects, but with shorter observations, leading to a larger value for  $\sigma_{\text{pip}}$ . The same panel shows an increase in variance

around 4400 Å, caused by the collimator mirror reflectivity (Guy et al., 2023). For reference, the distribution of fluxes is centered around  $5 \cdot 10^{-18} \text{erg s}^{-1} \text{cm}^{-2} \text{Å}^{-1}$  in the blue end of the spectra, and  $5 \cdot 10^{-19} \text{erg s}^{-1} \text{cm}^{-2} \text{Å}^{-1}$  in the red end.

To match the accuracy of our most reliable mock data (Herrera et al., 2023) and due to the limited amount of Lyman- $\alpha$  forest data available at longer wavelengths, we have limited our analysis to fluctuations fulfilling  $z < 3.79$ . Given that Lyman- $\alpha$  fluctuations can be related to a specific location in the wavelength grid ( $z = \lambda/\lambda_{\text{Ly}\alpha} - 1$ ), this redshift cut corresponds to a maximum wavelength of 5772 Å entering our analysis.

## 6.2.2 Description of the quasar catalog

Objects in the spectroscopic data are identified using the template-fitting code Redrock (Bailey et al. (2023), and for the special case of quasars, Brodzeller et al. (2023)). Redrock employs a set of templates as representatives of the main object classes observed by DESI: quasars, galaxies and stars. For each observed spectrum, Redrock determines the best-fitting redshift and template by comparing it to the set of templates. This process allows Redrock to accurately identify the objects in the DESI spectroscopic data.

However, Redrock sometimes misidentifies quasars as galaxies. To address this issue, a set of independent quasar-identification approaches, referred to as "afterburners" are also run. These afterburners can identify features missed by Redrock and improve the accuracy of quasar identification. In our case, the most relevant afterburners are QuasarNet (Busca and Balland, 2018; Farr et al., 2020a), the Mg II afterburner, and SQUEzE (Pérez-Ràfols et al., 2020). While only the first two methods were used to build the final catalog, all of them were tested during the SV phase (Alexander et al. (2023), also see Lan et al. (2023) for SV results on galaxies).

For the case of QuasarNET, its main role is to correct for cases where Redrock identifies an object as a low redshift galaxy when it is actually a high redshift quasar. In these cases, QuasarNet is used to re-identify the object as a quasar based on its spectral features using Machine Learning techniques and trained using visually inspected datasets. If QuasarNet identifies an object as a high redshift quasar, Redrock is run again with a high redshift prior. This will likely change the object identification to a high redshift quasar, and if confirmed, the object would be included in the final catalog.

The other relevant afterburner is the Mg II afterburner. Its main role is to correct for cases where Redrock misidentifies a quasar as a galaxy. For every object



identified as a galaxy by Redrock, the Mg II afterburner checks the width of the Mg II emission line. If the line is wide enough, the object is re-identified as a quasar and included in the final catalog.

For this release, only objects targeted and confirmed as quasars are used. The resulting catalogs include a total of 68 750 quasars for the EDR+M2 sample and 318 691 quasars for the EDR+M2 (see Table 6.1). This value includes all quasars in the input sample, being the actual value of quasars used for each region detailed in Table 6.2. The redshift and spatial distributions of the quasars is shown in Fig. 6.2, compared to the larger catalog used in eBOSS DR16 analysis (du Mas des Bourboux et al., 2020). The number of quasars in the EDR+M2 is clearly dominated by the first two months of main survey (M2), containing almost half of the objects in the eBOSS sample.

### 6.2.3 BAL and DLA information

Damped Lyman- $\alpha$  Absorption (DLA) systems caused by H I-rich galaxies affect the Lyman- $\alpha$  forest by generating absorption features that can interfere with the continuum fitting process. Damped Lyman- $\alpha$  Absorption (DLA) information is provided using a DLA finder based on a convolutional neural network and a gaussian process, for a full description of the DLA catalog see Zou et al. (2023). We selected from the full sample the objects where both the convolutional neural network and the gaussian process detected a DLA with a confidence larger than 50%.

Apart from DLAs, the Lyman- $\alpha$  forest can also be affected by broad absorption lines believed to be caused by the existence of ionized plasma outflows from the accretion disk. These BAL quasars can be added to the analysis, although they have to be appropriately masked (see Section 6.3.1). The algorithm for BAL identification was presented in Guo and Martini (2019), and the detailed description of the BAL catalog for DESI data is presented in Filbert et al. (2023).

The number of DLAs and BALs for both EDR+M2 and EDR+M2 samples is shown in Table 6.1. Given the better signal-to-noise in the EDR+M2 sample, the identification of DLA and BAL objects in this sample is higher in this smaller dataset.

## 6.3 Spectral reduction

The continuum fitting procedure estimates the expected flux for each of the quasars in the catalog, this process is essential for computing Lyman- $\alpha$  fluctuations. The



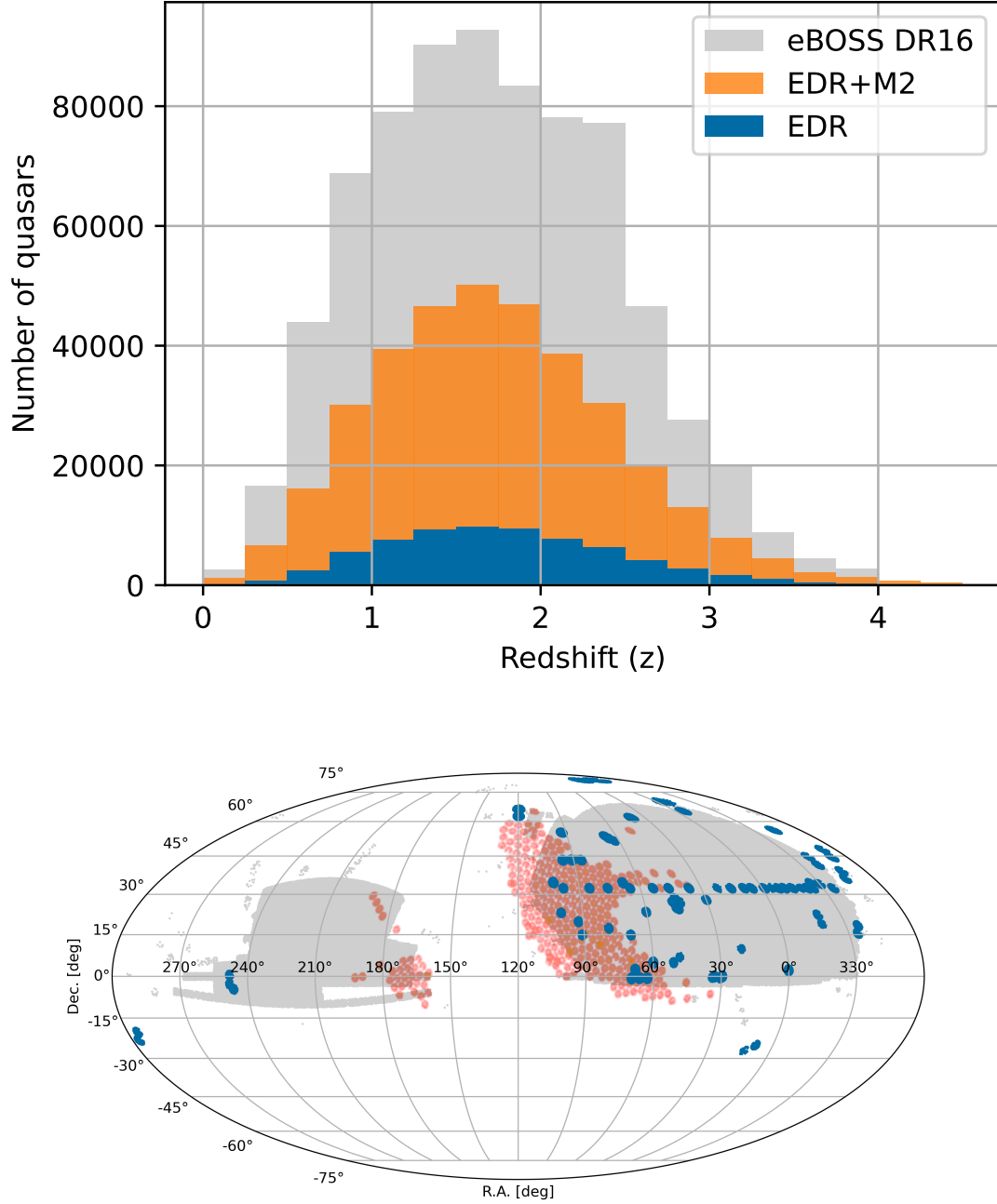


Figure 6.2: Distribution of objects in the two samples used in this publication, compared with the final eBOSS DR16 sample presented in (du Mas des Bourboux et al., 2020). Top: Redshift distribution of the three catalogs. Bottom: Sky distribution of the same catalogs. We observe that M2 makes a significant portion of the EDR+M2 sample. When compared to the eBOSS data, the EDR+M2 sample is approximately halfway to matching the number of objects in the former.

	EDR	EDR+M2
Quasars	68 750	318 691
DLAs	5 006	17 375
BALs	9 294	28 185

Table 6.1: Number of Lyman- $\alpha$  quasars in the two samples, number of quasars showing BAL features and number of DLAs affecting forests from the Lyman- $\alpha$  quasars. The better signal-to-noise in the EDR sample allows for the detection of a higher number of DLA and BAL features.

flux-transmission field can be defined as:

$$\delta_q(\lambda) \equiv \frac{f_q(\lambda)}{\bar{F}(\lambda)C_q(\lambda)} - 1, \quad (6.1)$$

where  $\bar{F}$  is the mean transmitted flux at a specific wavelength,  $C_q$  the unknown unabsorbed quasar continuum for quasar  $q$  and  $f_q$  its observed flux. The combination  $\bar{F}C_q(\lambda)$  is the mean expected flux of the quasars, and is the quantity that we fit for the spectra.

Continuum fitting is the procedure to compute the flux-transmission field. It can be split in an initial clean-up phase and a second phase where the quasar continuum is actually fitted. The clean-up phase involves two sequential procedures: masking multiple unmodelled features and re-calibrating the spectra to eliminate residuals from the DESI calibration procedure.

In this Section we first explain these two procedures, and then provide details on the core of the continuum fitting process. We will end by describing the differences between our analysis and previous SDSS methods.

### 6.3.1 Masks (sky lines, galactic absorption, DLA, BAL)

There are multiple features in the spectra that are not caused by Lyman- $\alpha$  fluctuations but are still present in our spectra due to contamination. In order to simplify the cosmological modeling when using Lyman- $\alpha$  fluctuations for correlation measurements, we mask these features and remove them from our analysis.

The three type of masks applied in our analysis are: DESI pipeline; BAL and DLA, applied to absorption features in the forest<sup>2</sup> region; and galactic absorption and sky emission lines that appear in the spectrograph when observing Lyman- $\alpha$  quasars. We note that the DESI pipeline mask is applied before the individual observations are co-added (see Section 6.2.1). The other masks are applied after the coaddition.

<sup>2</sup>We will refer to each individual line-of-sight as forest, following the convention from previous Lyman- $\alpha$  publications.

### 6.3.1.1 DESI pipeline masking

A masking process is applied within the DESI pipeline to identify bad pixels in the spectra. These are typically caused by CCD defects or cosmic rays hitting the spectrograph CCD. This mask is found to only affect about 0.1% of the DESI pixels used for the Lyman- $\alpha$  analysis. Given the small fraction of pixels discarded, we decided to remove these pixels from the analysis.

### 6.3.1.2 DLA and BAL masking

DLA absorption affects our spectra by imprinting itself in the spectra of quasars, resulting in zero flux around the true redshift of the DLA and damping wings further away. A secondary effect of this is the reduction in the mean flux of the affected spectra, affecting the overall mean absorption. This biases our estimate of  $\bar{F}C_q$ , potentially affecting all the quasars in the sample (see Section 6.3.3).

Although it is possible to include the absorption features from DLAs into our models for the correlation functions, for simplicity, and following the prescription in previous analyses (du Mas des Bourboux et al., 2020), we mask the regions of the Lyman- $\alpha$  forest that are affected by identified DLAs when the DLA reduces the transmission by more than 20%. We then correct the absorption in the wings using a Voigt profile as suggested by Noterdaeme et al. (2012), being able to include the pixels affected by these wings without affecting the mean absorption.

The top panel of Fig. 6.3, shows the fraction of pixels that have been masked due to DLAs as a function of observed wavelength. Although there is an increase in the number of detected DLAs with redshift, the number of masked pixels is always smaller than 5%.

Regarding BALs, in previous analyses quasars exhibiting BAL features were directly excluded from the analysis, although they were later used as tracers for cross-correlations between quasars and the Lyman- $\alpha$  forest. However, Ennesser et al. (2022) showed that quasars with BAL features could be safely included in the analysis if all expected locations of these absorption features are masked. Here, we follow their proposed approach.

In the bottom panel of Fig. 6.3, we observe the fraction of pixels masked due to BALs as a function of rest-frame wavelength. In this case, we can see a pattern of absorption that matches the expected absorption features described in Ennesser et al. (2022). Since not all the BAL quasars suffer from the same absorption profile, we expect the BAL-masked pixel fraction to be maximal around the central absorption and decrease as we go away from it. It is worth noting that for the

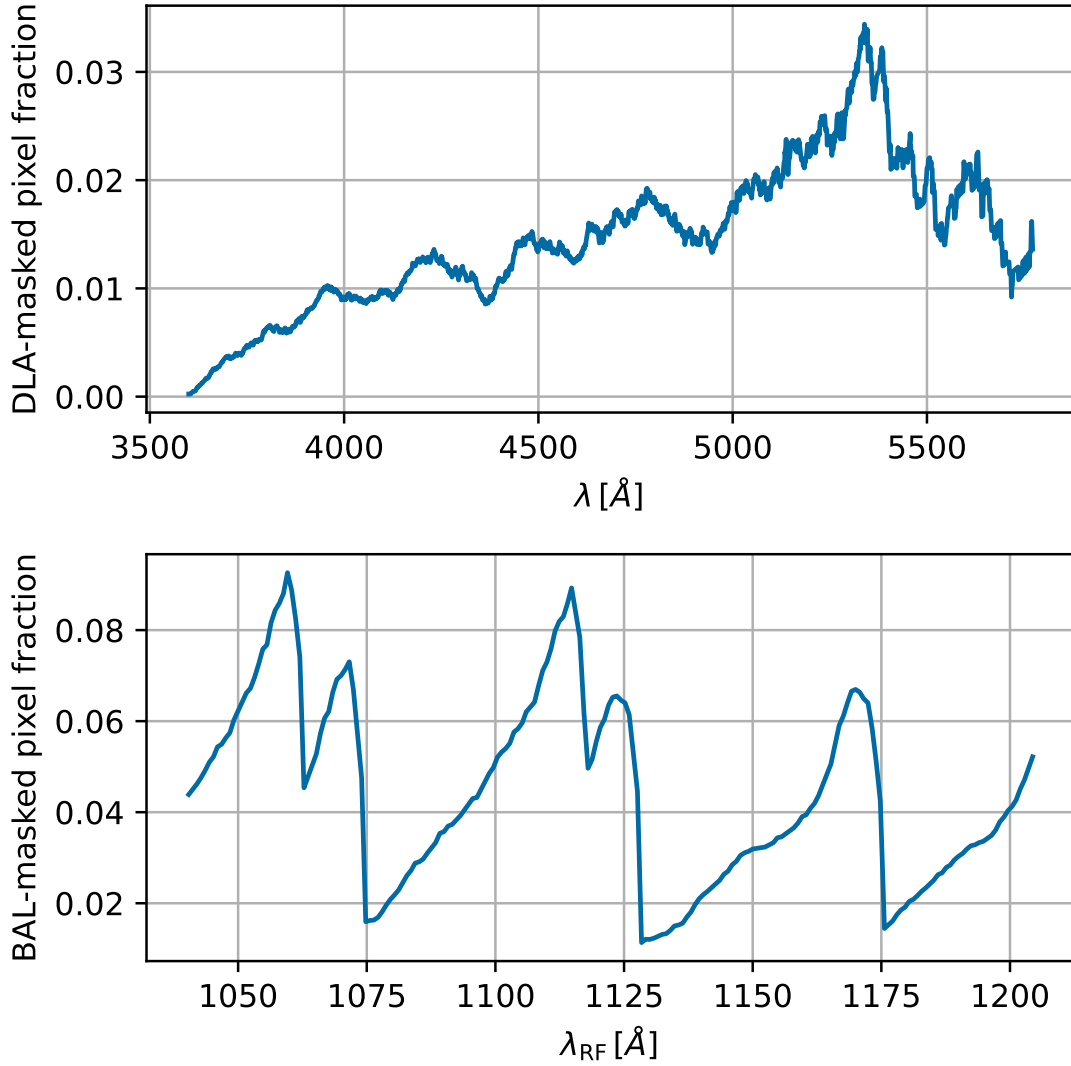


Figure 6.3: Fraction of pixels masked due to DLAs (top) and BAL (bottom) features. As expected, the number of detected DLAs increase with redshift (and therefore with  $\lambda$ ), yielding a fraction of masked pixels always below the 5%. For the BAL case, we show the masked fraction as a function of  $\lambda_{\text{RF}}$ , which allow us to observe the strong wavelength dependence of the masking, associated with emission lines for different elements. This Figure used the full EDR+M2 dataset.

quasars used in the Lyman- $\alpha$  region continuum fitting, the percentage of BAL quasars is 16% for the EDR+M2 sample and 23% for the EDR+M2 sample.

### 6.3.1.3 Galactic absorption and sky emission masking

There are certain absorption and emission features in our spectra that differ from Lyman- $\alpha$  fluctuations or other absorbers in the Intergalactic Medium (IGM). These

absorption and emission features can be easily identified because they affect specific wavelengths, resulting in sharp features when the spectra of multiple quasars is combined. We can remove them by simply masking out the corresponding wavelengths from our analysis.

The two most significant features in this regard are galactic absorption and sky emission. Galactic absorption is caused by material in the Interstellar Medium (ISM) absorbing at specific wavelengths. The ISM absorption for some of these lines cannot be easily separated from the intrinsic absorption in the atmospheres of the stars used for flux calibration, and thus they are not properly accounted for. Here, we are affected by the Ca K and H transitions, and we mask the corresponding wavelengths (see Fig. 6.4).

Sky emission comprises emission lines generated by atmospheric effects and are mostly corrected in the modelling of sky lines. However, inaccuracies in this modelling result in spurious features in our spectra. Here, we also mask the affected wavelengths (see Fig. 6.4).

In Fig. 6.4, we present the measurement of the estimated  $\overline{1 + \delta_q(\lambda)}$  in the C III region (see Table 6.2). Two kind of features can be observed in this plot: first, we observe the sharp features corresponding to the mentioned galactic absorption and sky emission; then we observe smooth features caused by inaccuracies in the DESI calibration process. To correct for the former, we mask the pixels in the wavelength intervals shown in the plot. The later can be corrected through re-calibration (see Section 6.3.2).

Here we use the C III region to build our masks as it lacks the Lyman- $\alpha$  absorption features. This makes it easier to estimate absorption and emission effects.

### 6.3.2 Re-calibration

DESI calibrates fluxes using standard stars (Guy et al., 2023). However, inaccuracies in the modelling of these calibration stars introduce features in the measured spectra when fluxes are predicted for other objects (such as quasars). These features can be observed by examining the mean  $\delta_q(\lambda)$  in any region of the spectra, particularly in the regions without Lyman- $\alpha$  absorption, where the variance of the measured flux is expected to be lower. This is possible because the features caused by calibration defects are function of observed wavelength, while the continuum estimate is a function of rest-frame wavelength. Looking again at Fig. 6.4, we can observe smooth features in the stack of fluctuations  $\overline{(1 + \delta_q(\lambda))}$  alongside the masked sharp features.

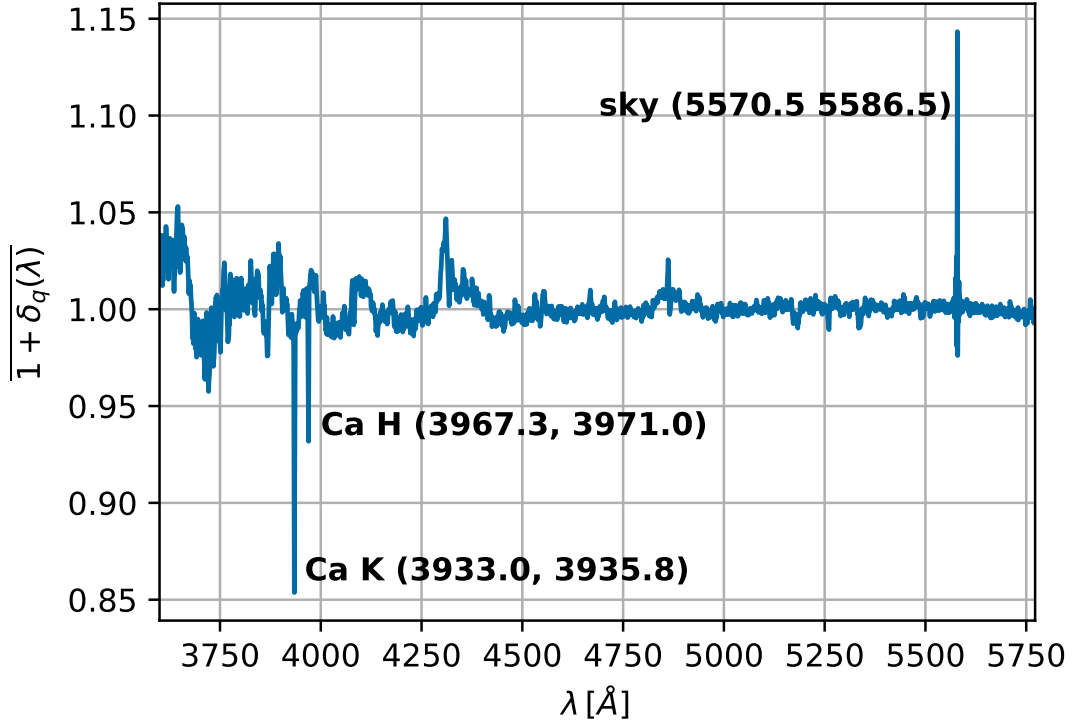


Figure 6.4: Weighted average of the flux-transmission field measured in the C III region. This measurement has been performed without masking sharp features, and they can be clearly observed at different positions in the spectrograph. The three masked regions are specified in the plot, showing the wavelength range affected by the mask. Smooth features in this measurement can also be observed. Their impact can be corrected through the flux re-calibration (see Section 6.3.2). For this Figure we used the full EDR+M2 dataset.

These features are correlated between different forests, and could therefore bias the measurement of the correlation function. To avoid this, on top of the DESI pipeline calibration, we re-calibrate our spectra. As stated above, these features affect the same region of observed wavelength regardless of the rest-frame position where they act. Therefore, one could in principle use the entire quasar spectra for calibration purposes. However, the larger spectral diversity near quasar emission lines can unnecessarily complicate this. In consequence, we search for a featureless region redwards of the Lyman- $\alpha$  emission line. We will refer to all the candidate regions as re-calibration regions.

In the absence of Lyman- $\alpha$  fluctuations, flux fluctuations are ideally only caused by noise. In any of the re-calibration regions, the measurement of  $\overline{1 + \delta_q(\lambda)}$  is expected to be consistent with 1; and its measurement can be considered a null test that is not fulfilled in general given the issues described above.

Region	$\lambda_{\text{RF}, \text{min}}$ Å	$\lambda_{\text{RF}, \text{max}}$ Å	# forests EDR	# forests EDR+M2
C III	1600	1850	49 810	233 310
C IV	1410	1520	41 445	189 984
Mg II-R	2900	3120	7 290	33 936
Mg II	2600	2760	13 373	62 628
S IV	1260	1375	34 044	152 979
Lyman- $\alpha$	1040	1205	20 281	88 511

Table 6.2: Statistics for the regions considered during the analysis, the region span is defined in the quasar rest-frame wavelength ( $\lambda_{\text{RF}}$ ). The number of forests corresponds to the number of quasars whose spectra can be observed in the spectrograph. A minor number of forests are rejected due to low signal-to-noise ratio (SNR) or due to them being too short.

In the bottom panel of Fig. 6.5, we show the measurement of  $\overline{1 + \delta_q(\lambda)}$  for multiple different candidate re-calibration regions. The fact that all of them show similar features, in spite of being at different regions of the spectra, suggests that all fluxes can be corrected consistently. Furthermore, we see comparable features in the White Dwarf average residuals in the top panel of the same Figure. This similarity can be seen in some of the regions of the spectrum, especially for  $\lambda \in [3600, 4200]$  Å, and it further justifies the re-calibration process.

In previous analyses, a region to the right of the Mg II emission line (Mg II-R) was selected to re-calibrate fluxes, partly because it is located further to the right of the spectra, reducing potential contamination from other absorption lines. However, in this work we selected the C III region ( $\lambda \in [1600, 1850]$  Å) due to the larger number of pixels available and given the similar behavior compared to the other regions. Fig. 6.6 shows the number of pixels at each wavelength bin of size  $\Delta\lambda = 55.58$  Å for the different regions, and the number of forests available for each region can be found in Table 6.2. In both cases, we see that C III has the largest number of pixels available for the analysis.

The choice of a re-calibration region at a larger wavelength than the Lyman- $\alpha$  forest also leads to an increase in the number of quasars that can be used for this process. The Lyman- $\alpha$  forest analysis includes quasars with redshifts in the range  $z \in (2.1, 3.7)$ , while for the C III region, quasars in the range  $z \in (0.9, 2.6)$  are included. By looking at the quasar distribution (Fig. 6.2), we see that the number of objects in the second range will be larger, and hence a larger number of pixels available in the re-calibration region.

We use a similar procedure as du Mas des Bourboux et al. (2020) for this re-calibration, correcting all the fluxes by using the mean flux in the re-calibration

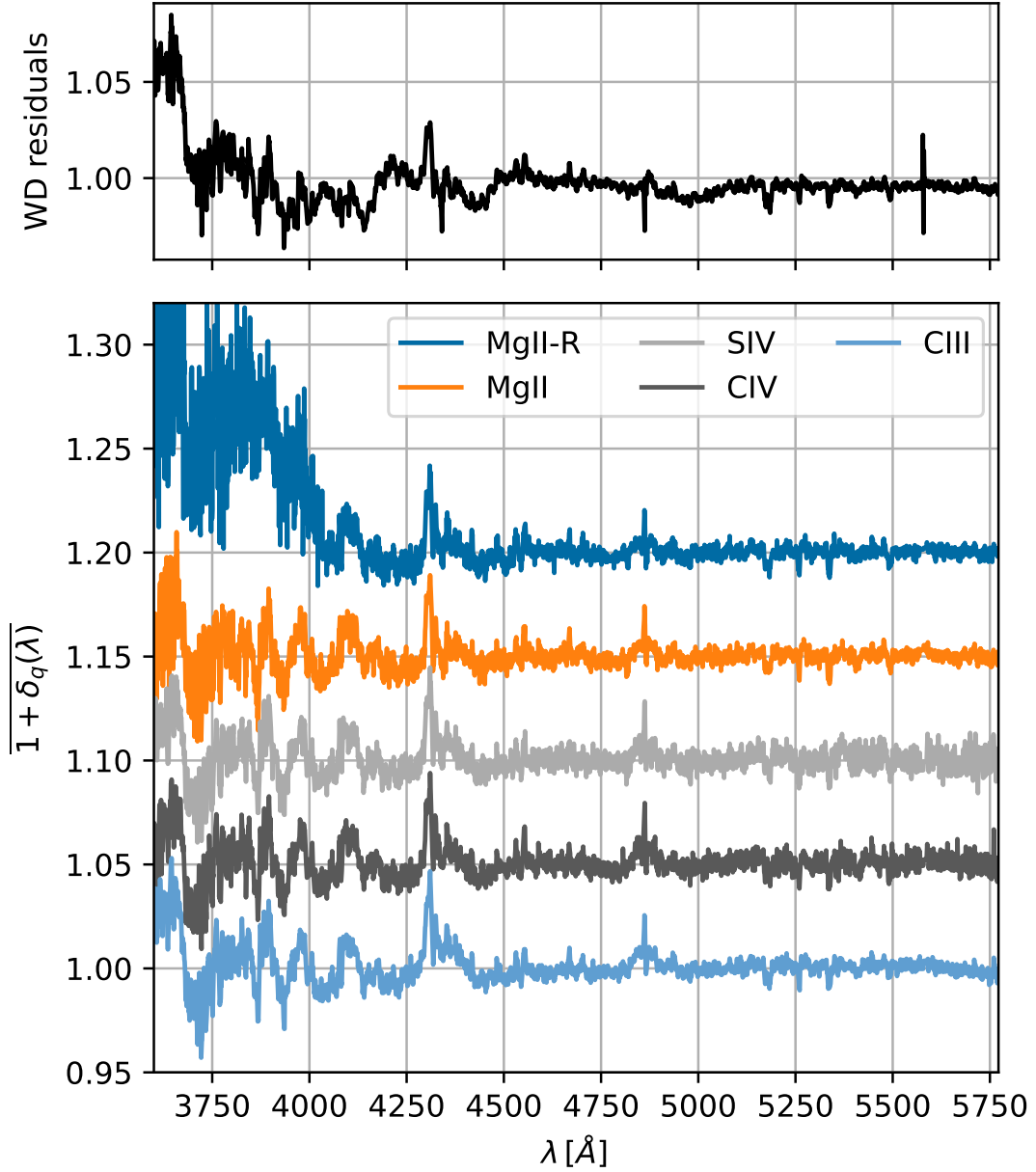


Figure 6.5: Bottom: Weighted average of the flux-transmission field measured at different regions of the spectra. Its value has been shifted to better distinguish the different regions. As opposed to the results shown in Fig. 6.4, sharp features are not present here because these samples have already been masked. The smooth features in the spectra are similar for all the measured regions, being the Mg II-R an outlier in this tendency, likely to be caused due to the reduced number of pixels available for this region at low wavelengths (see Fig. 6.6). Results are computed from the full EDR+M2 sample. Top: Average residuals to White Dwarf (WD) spectra in the blue arm of the DESI spectra, as seen in Guy et al. (2023). A similar trend can also be observed here between these residuals and our measured average of the flux-transmission field, especially at smaller wavelengths, further justifying our re-calibration process.



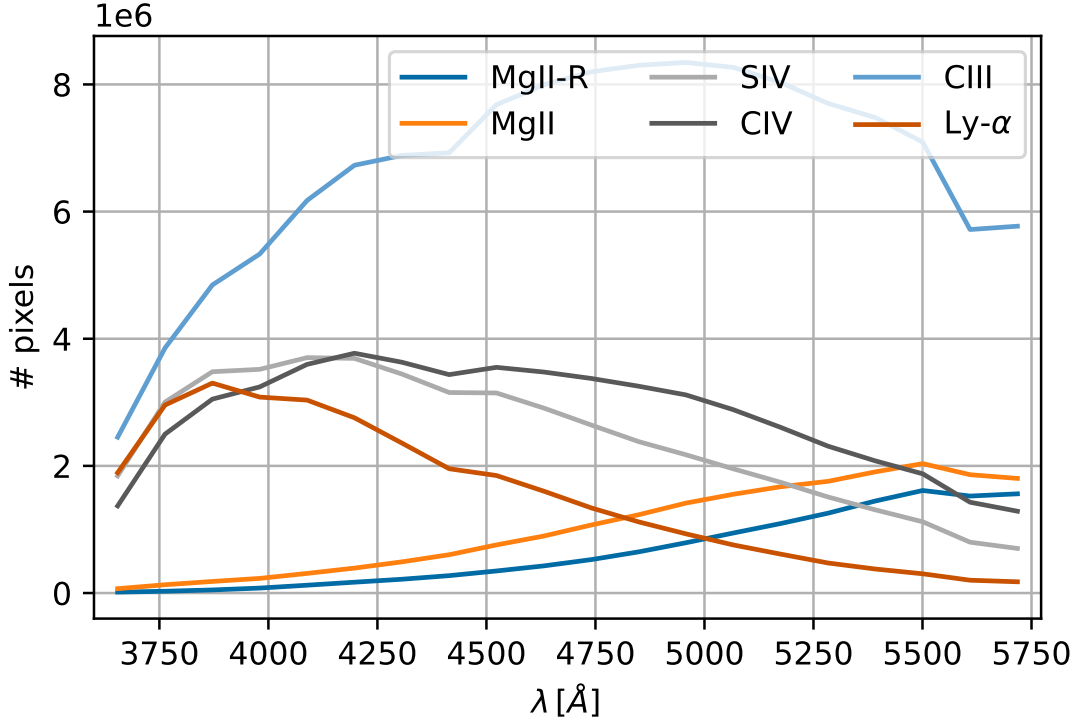


Figure 6.6: Number of pixels available for the different regions measured in bins of  $\Delta\lambda = 55.58 \text{ \AA}$ . Given the location of the different regions at different positions in the quasar spectra, the number of available pixels is different for each of them. This is because our spectral coverage lies in the range  $(3600, 5772) \text{ \AA}$ . Thanks to the larger number of pixels available for the C III region, it was selected as the region to be used for the re-calibration process. This Figure used the full EDR+M2 sample.

region:

$$f_{\text{new}}(\lambda_i) = f_{\text{orig}}(\lambda_i) / \overline{f_{\text{calib, orig}}(\lambda_i)} . \quad (6.2)$$

The error reported by the pipeline for these fluxes also needs to be corrected through:

$$\sigma_{\text{pip, new}}(\lambda_i) = \sigma_{\text{pip, orig}}(\lambda_i) / \overline{f_{\text{calib, orig}}} . \quad (6.3)$$

After this correction, smooth features in the spectra will be alleviated, which prevents biased results. In previous analyses, a second correction was applied to correct the estimation of the flux variance provided by the pipeline (see Section 6.3.4). However, the better performance of DESI in this aspect allows us to skip this step. Evidence in this direction is the better behavior of the estimated correction to the pipeline error (see Fig. 6.7).

### 6.3.3 Continuum fitting

The core of the continuum fitting process consists of obtaining the expected flux  $\bar{F}C_q$  for each quasar in the sample. This allows for the determination of the flux-transmission field (Eq. (6.1)). The process is performed iteratively, with several quantities fitted simultaneously, and performed on each of the quasar regions independently: in the case of a re-calibrated Lyman- $\alpha$  analysis, it is firstly performed in the re-calibration region ( $C_{\text{III}}$  in this case), and afterwards in the Lyman- $\alpha$  region.

In order to simplify the process, the expected flux  $\bar{F}(\lambda)C_q(\lambda)$  is assumed to be a universal function of rest-frame wavelength,  $\bar{C}(\lambda_{\text{RF}})$ , corrected by a first degree polynomial in  $\log \lambda$ :

$$\bar{F}(\lambda)C_q(\lambda) = \bar{C}(\lambda_{\text{RF}}) \left( a_q + b_q \frac{\Lambda - \Lambda_{\min}}{\Lambda_{\max} - \Lambda_{\min}} \right), \quad (6.4)$$

where  $\Lambda \equiv \log \lambda$  and  $\Lambda_{\min, \max}$  identify its minimum and maximum values inside the region to be fitted. That is, the spectra of quasars in our sample are assumed to have the same underlying shape or continuum for all objects, allowing for variations in the amplitude and tilt. In the analysis, the transmitted flux  $\bar{F}$  and the quasar continuum  $C_q$  cannot be fitted independently, although they are not needed separately in our analysis. For this reason, we choose to directly estimate the expected flux of quasars.

The parameters  $(a_q, b_q)$  are fitted by maximizing the likelihood function

$$2 \ln L = - \sum_i \frac{\left[ f_i - \bar{F}C_q(\lambda_i, a_q, b_q) \right]^2}{\sigma_q^2(\lambda_i)} - \sum_i \ln \left[ \sigma_q^2(\lambda_i) \right], \quad (6.5)$$

where  $\sigma_q^2(\lambda)$  is the variance of the flux  $f_i$ . This value has to be estimated, and since it depends on  $(a_q, b_q)$ , we include this dependence in the likelihood function.

The full variance of the flux,  $\sigma_q^2(\lambda)$ , includes not only the obvious contribution from the noise estimated by the DESI pipeline but also the intrinsic variance of the Lyman- $\alpha$  forest<sup>3</sup>. We account for the intrinsic variance in the following way:

$$\frac{\sigma_q^2(\lambda)}{\left( \bar{F}C_q(\lambda) \right)^2} = \eta(\lambda) \tilde{\sigma}_{\text{pip}, q}^2(\lambda) + \sigma_{\text{LSS}}^2(\lambda). \quad (6.6)$$

where  $\sigma_{\text{LSS}}^2$  is the mentioned intrinsic variance of the Lyman- $\alpha$  forest, and  $\tilde{\sigma}_{\text{pip}, q} = \sigma_{\text{pip}, q}(\lambda) / \bar{F}C_q(\lambda)$  where  $\sigma_{\text{pip}, q}$  is flux variance as estimated by the DESI pipeline.

<sup>3</sup>This quantity is expected to be very small outside the Lyman- $\alpha$  forest region.

In this expression we also include a correction  $\eta(\lambda)$  to account for inaccuracies in the pipeline noise estimation. We discard here an extra term in this expression accounting for quasar variability at high SNR, which was previously used in du Mas des Bourboux et al. (2020) (see Section 6.3.4.2 for details). It is worth noting that  $\sigma_q^2$  here is the variance of the flux and therefore has flux units, whether all quantities at the right hand side are dimensionless.

In the iterative continuum fitting process, we estimate the quantities  $\eta$ ,  $\sigma_{\text{LSS}}^2$ ,  $a_q$ ,  $b_q$ ,  $\bar{C}(\lambda_{\text{RF}})$ . The sequential steps of the process are:

1. **Assume wavelength-independent values for unknown quantities:** A flat assumption is assumed for all the quantities to be fitted and measured. In practice, this means setting  $\bar{C}(\lambda_{\text{RF}}) = 1$ ,  $\eta(\lambda) = 1$  and  $\sigma_{\text{LSS}}^2(\lambda) = 0.1$ .
2. **Fit the per-quasar parameters ( $a_q, b_q$ ):** Using the current values of  $\bar{C}(\lambda_{\text{RF}})$ ,  $\eta(\lambda)$  and  $\sigma_{\text{LSS}}^2(\lambda)$ , the pair of values ( $a_q, b_q$ ) is estimated for each individual forest through the minimization of the likelihood defined in Eq. (6.5).
3. **Estimate the flux-transmission field  $\delta_q(\lambda)$ :** Using the current value  $\bar{C}(\lambda_{\text{RF}})$  and the best-fit parameters for ( $a_q, b_q$ ), we compute the expected flux for each quasar following Eq. (6.4). This allows for the computation of the fluctuations around the expected flux ( $\delta_q(\lambda)$ ) as defined in Eq. (6.1).
4. **Fit the variance functions:** The functions  $\eta(\lambda)$  and  $\sigma_{\text{LSS}}^2(\lambda)$ , defined in Eq. (6.6), are now fitted using the estimated  $\delta_q(\lambda)$  from the previous step. In order to do so, different values of the flux-transmission field are grouped by wavelength and  $\tilde{\sigma}_{\text{pip}}$ . We take 20 bins for the wavelength split and 100 for the split on  $\tilde{\sigma}_{\text{pip}}$ , generating a grid of 2000 points. We compute the variance  $\sigma^2(\lambda, \tilde{\sigma}_{\text{pip}}) = \overline{\delta_q^2(\lambda, \tilde{\sigma}_{\text{pip}})}$  for each point in the grid<sup>4</sup>. The functions  $\eta(\lambda)$  and  $\sigma_{\text{LSS}}^2$  are fitted for each of the 20 wavelengths independently using the following likelihood function:

$$2 \ln L = - \sum_{\tilde{\sigma}_{\text{pip}}} \frac{\left[ \overline{\delta_q^2(\lambda, \tilde{\sigma}_{\text{pip}})} - \eta(\lambda) \tilde{\sigma}_{\text{pip}}^2 - \sigma_{\text{LSS}}^2(\lambda) \right]^2}{\overline{\delta_q^4(\lambda, \tilde{\sigma}_{\text{pip}})}}, \quad (6.7)$$

where the sum in  $\tilde{\sigma}_{\text{pip}}$  include all the valid bins in  $\tilde{\sigma}_{\text{pip}}$  (at most 100). The points in the grid computed using less than 100 pixels are considered unreliable and discarded from the fits. We note that in those cases where the fit does not converge (for example, if there are too few quasars), then the default values from step Item 2 are kept.

---

<sup>4</sup>Here we assume  $\overline{\delta_q(\lambda, \tilde{\sigma}_{\text{pip}})} = 0$  as an approximation. Fluctuations around this assumption are small (see Fig. 6.8)

5. **Recompute the mean expected flux:** At this stage we update the value of  $\overline{C}(\lambda_{\text{RF}})$ . This is performed by computing the weighted average of all quasars expected flux sharing the same  $\lambda_{\text{RF}}$  value. Here, we use the optimal weights as defined in Eq. (6.10) (see Section 6.4.1).
6. **Compute and save relevant statistics:** Relevant statistics are computed at each iteration and stored in `delta_attributes_iteration{i}.fits.gz` files (where  $i$  is the iteration number). The saved statistics include the stack of fluctuations  $\overline{(1 + \delta_q(\lambda))}$ , the fitted variance functions  $(\eta, \sigma_{\text{LSS}}^2)$ , the mean continuum  $\overline{C}(\lambda_{\text{RF}})$  and fit metadata including the tilt and slope values  $(a_q, b_q)$ , the number of pixels used for the fit and the  $\chi^2$  of the fit.
7. **Continue next iteration starting in step Item 2:** The next iteration starts using the updated values of  $\overline{C}(\lambda_{\text{RF}})$ ,  $\eta$  and  $\sigma_{\text{LSS}}^2$ .

This process is performed 5 times, resulting in stable estimates of all the defined quantities.

In Fig. 6.7, we show the final estimates for the two fitted function  $\eta$  and  $\sigma_{\text{LSS}}^2$  for the two regions considered in this analysis (Lyman- $\alpha$  and C III) and for the two different samples (EDR and EDR+M2). The correction to the pipeline estimated variance,  $\eta$ , shows a  $\sim 2\%$  deviation from the ideal value of 1 in the case of EDR+M2 and up to  $\sim 7.5\%$  deviation in the case of EDR. This difference is due to the worse estimation of the pipeline-reported variance for the first part of the SV program (see Ravoux et al. (2023) for details). For  $\sigma_{\text{LSS}}^2$ , its estimation is similar for both samples, showing a clear dependence on wavelength (and hence redshift) for the Lyman- $\alpha$  region and an expected (due to the lack of Lyman- $\alpha$  fluctuations) nearly zero value for the C III region. The fit fails at the higher wavelength values in the case of the EDR sample due to the small number of pixels available at these wavelengths, and then the initial value is kept.

Fig. 6.8 shows the mean flux-transmitted field for the same two regions and samples. In this case, differences between the two samples are smaller, only showing a higher variance in the case of the smaller (EDR) sample. The variances are particularly worse at the largest wavelengths in the Lyman- $\alpha$  region due to the reduced number of pixels available. The C III region, for which no re-calibration has been performed, still shows the smooth features already discussed in Section 6.3.2. They do not appear in the already re-calibrated Lyman- $\alpha$  region, although there is a higher variance caused by the less number of pixels (as shown in Fig. 6.6).

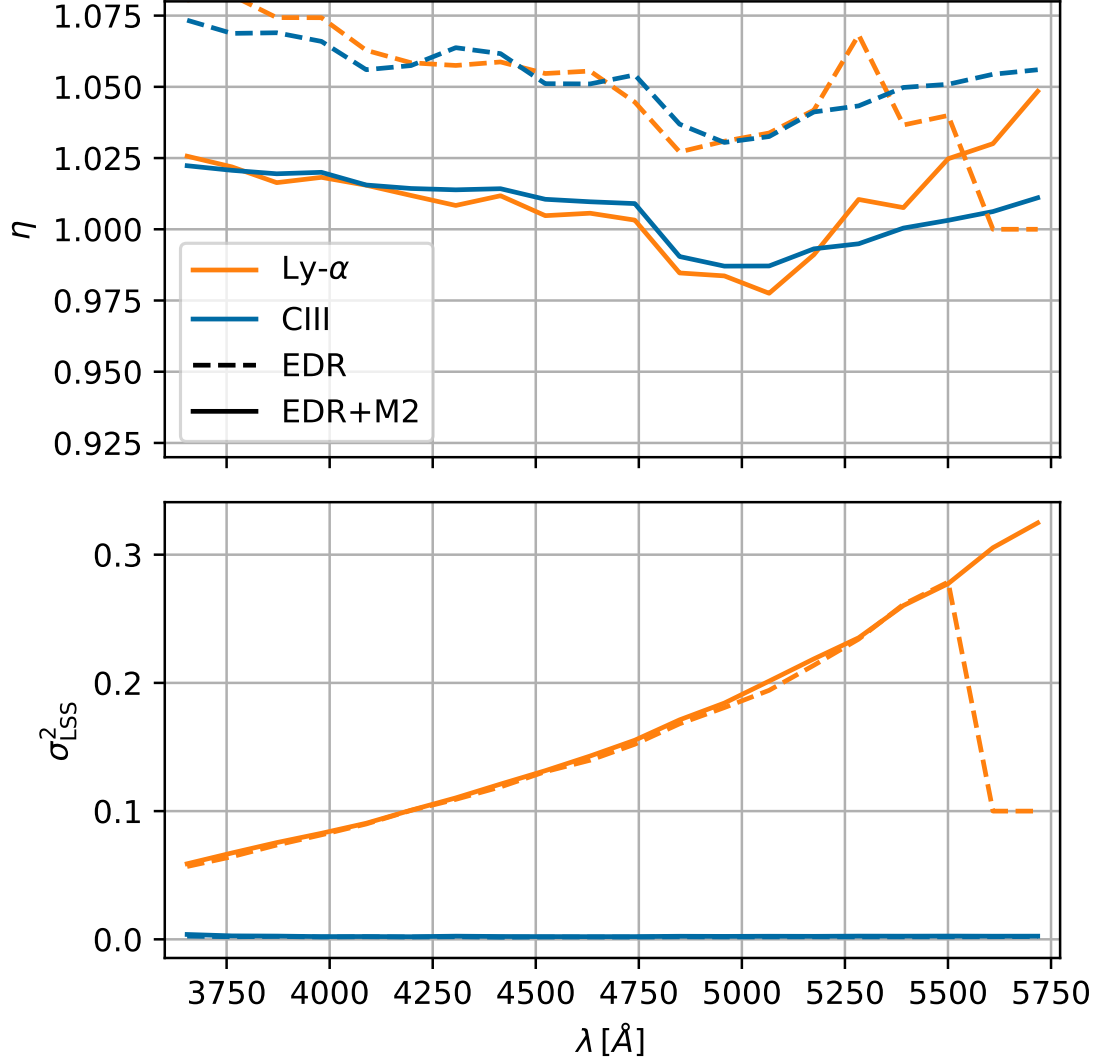


Figure 6.7: Wavelength evolution of the fitted parameters  $\eta$  and  $\sigma^2_{\text{LSS}}$ , measured in both the C III and Lyman- $\alpha$  regions for the EDR (dashed) and EDR+M2 (solid) samples. Top: The pipeline error correction  $\eta$  is found to be larger for the EDR sample, caused by the worse estimation of the pipeline-reported variance for the first part of the SV program. Bottom:  $\sigma^2_{\text{LSS}}$ , in this case it is consistent between the two samples. As expected the C III region shows a value close to 0 for all wavelengths, while for the Lyman- $\alpha$  region follows the expected increase in its intrinsic variance with redshift. In both  $\eta$  and  $\sigma^2_{\text{LSS}}$  measurements for the EDR sample, the fitted parameters could not be obtained for the larger wavelength value due to the reduced number of pixels available, falling to the default values  $\eta = 0.1$  and  $\sigma^2_{\text{LSS}} = 0.1$ .

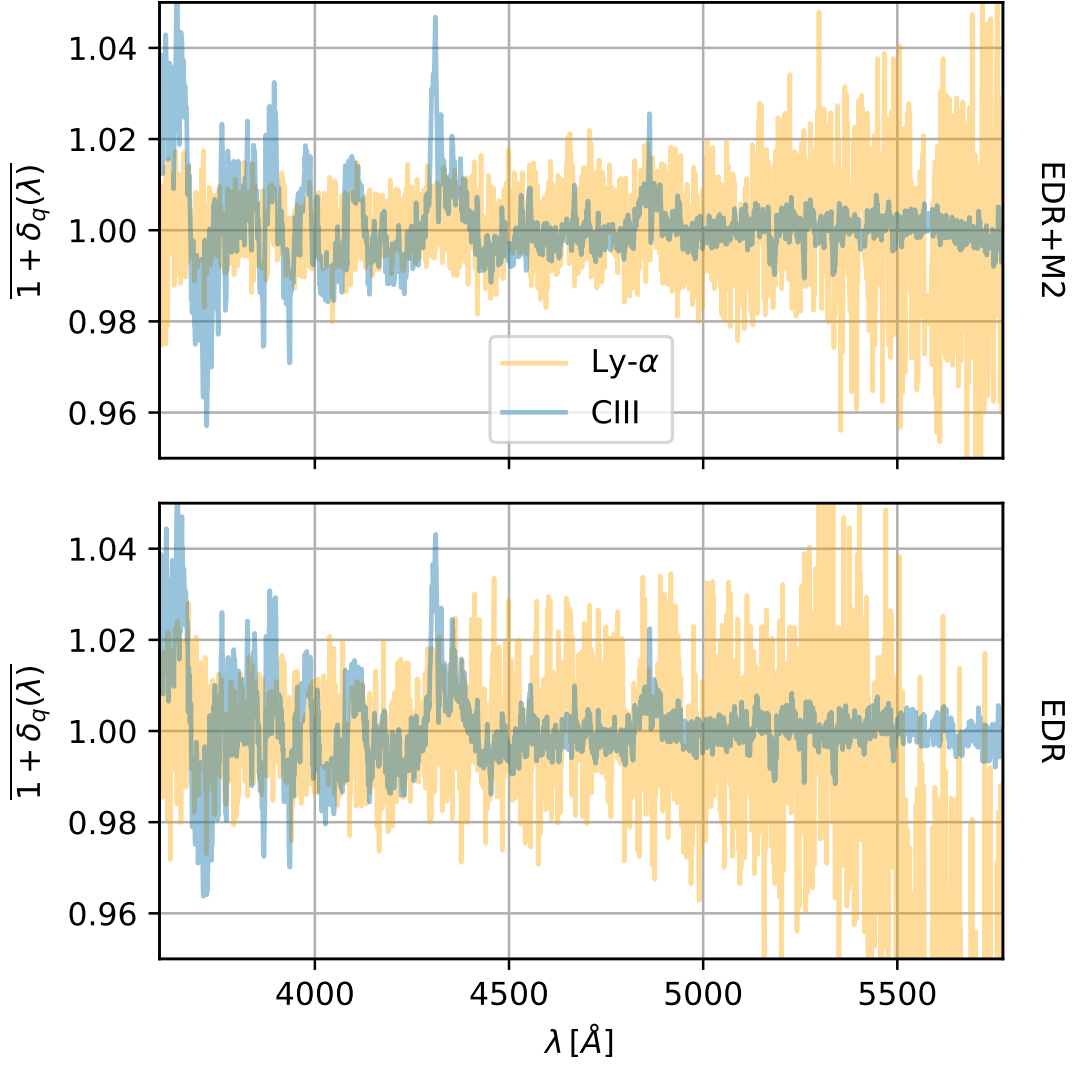


Figure 6.8: Measurement of the weighted mean of the flux-transmission field for both the Lyman- $\alpha$  and C III regions. The Lyman- $\alpha$  region shows an expected higher variance at all wavelengths compared to the C III region, although it lacks smooth features thanks to the re-calibration process. Similarly, the EDR sample shows a higher variance than EDR+M2. In both cases, this is caused by the decrease in number of pixels available. Given the low number of pixels available at larger wavelengths for the EDR sample for the Lyman- $\alpha$  region, it departs from the expected unity behavior.

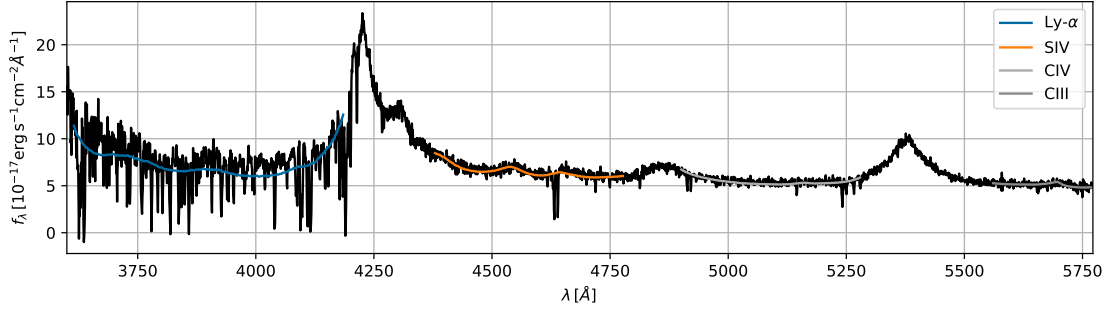


Figure 6.9: Example of a high signal-to-noise quasar spectrum. The mean expected flux  $\overline{C}(\lambda_{\text{RF}})$  as in Eq. (6.4) is shown for the Lyman- $\alpha$ , S IV, C IV and C III regions. The quasar has a redshift  $z_q = 2.495$  and is identified as a DESI object with TARGETID=39628443918272474. As in this example, we occasionally observe metal absorption in the calibration regions.

#### 6.3.3.1 Example of quasar fit

An example of a high SNR quasar spectrum is shown in Fig. 6.9. We can see the Lyman- $\alpha$  forest at the left side of the Lyman- $\alpha$  emission line ( $\lambda_{\text{RF}} = 1215.7 \text{ \AA}$ ), with the characteristic absorption features. The solid lines show the mean expected flux for that particular forest at different spectral regions. Metal absorption features can be observed in this example for the S IV and C IV re-calibration regions.

### 6.3.4 Changes with respect to previous analyses

In this work, we include<sup>1</sup> some changes compared to the last Lyman- $\alpha$  forest analysis conducted using SDSS data (du Mas des Bourboux et al., 2020). These changes were necessary due to the differences in the way data is structured in DESI as compared to SDSS. One of the most significant is the linear pixelization of the wavelength grid.

There are three relevant changes in the way the analysis is performed. First, the re-calibration has been updated to account for the improvements offered by the new instrument. Second, the variance estimation process has been simplified. Third, the re-calibration region has been switched to C III. We have already discussed the third change in Section 6.3.2, now we will describe the first two changes in detail.

#### 6.3.4.1 Re-calibration of spectra in SDSS

The improved estimation of pipeline noise ( $\sigma_{\text{pip}}$ ) in DESI has eliminated the need for multiple re-calibration steps as it was performed in the SDSS analysis.

For SDSS analyses, re-calibration was performed in two steps, both of which were conducted in the same re-calibration region. The first step was equivalent to the one used in this analysis (Section 6.3.2). After the first step, a second re-calibration step was added to further correct the reported variance of the pipeline. In this second step, the continuum fitting process was run again to obtain a new estimation of  $\eta$ .

Then, this new value of  $\eta$  is applied in the main Lyman- $\alpha$  fluctuations run to correct for the values of  $\sigma_{\text{pip}}$ :

$$\sigma_{\text{pip}}^2 (\text{Ly}\alpha)(\lambda) = \eta^{(\text{calib2})}(\lambda) \sigma_{\text{pip}}^2 (\text{calib2})(\lambda). \quad (6.8)$$

The extra re-calibration step in SDSS aimed to make the final estimation of  $\eta$  closer to 1, but it also added an unnecessary layer of complexity to the process with the sole purpose of doing so. However, it did not modify the product  $\eta \cdot \sigma_{\text{pip},q}^2$ , and as a result, the overall variance assigned to pipeline noise was unchanged.

#### 6.3.4.2 Variance estimator

In Eq. (6.6), we have defined the model for our variance of the flux  $\sigma_q^2(\lambda)$ . Its expression has been simplified from previous SDSS analysis, where this variance was modeled as:

$$\frac{\sigma_q^2(\lambda)}{(\bar{F}C_q(\lambda))^2} = \eta(\lambda)\tilde{\sigma}_{\text{pip},q}^2(\lambda) + \sigma_{\text{LSS}}^2(\lambda) + \frac{\epsilon(\lambda)}{\tilde{\sigma}_{\text{pip},q}^2(\lambda)}. \quad (6.9)$$

The additional term was added to account for the observed increase in variance at high SNR, which is likely caused by the diversity of quasar spectra.

In Fig. 6.10, we present the weight of each term in Eq. (6.9) contributing to the total variance, selecting only the highest SNR quasars. This shows that the effect of this added term is small, and for simplicity we decided to not include it in our analysis. The plot displays only the top 5 % higher SNR quasars since this subset is expected to have the largest contribution (the third term in Eq. (6.9) becomes larger). It is worth noting that the contribution is too small to be discernible when all the objects in the sample are considered.

However, it is important to take into account quasar diversity in future analyses. We plan to incorporate a term in Eq. (6.6) that considers the variance in the quasar's continuum. This term will be dependent on rest-frame wavelength,  $\sigma_C^2(\lambda_{\text{RF}})$ , and will also help us select the region of the quasar suitable for Lyman- $\alpha$  analyses, showing us how close the emission lines we can get.



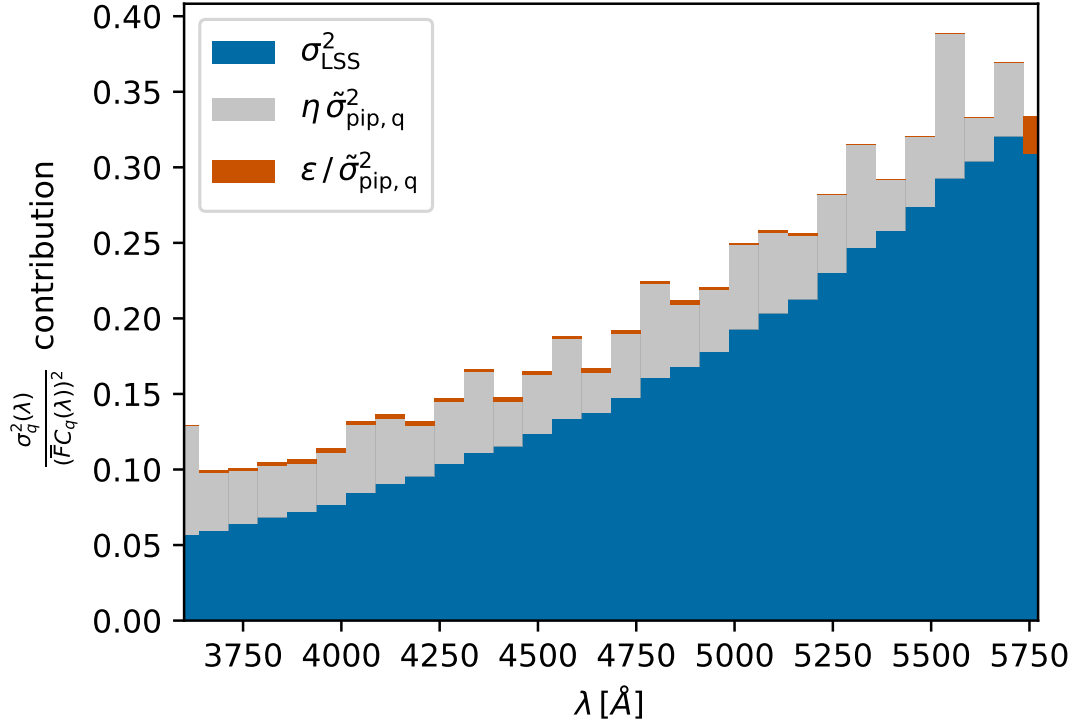


Figure 6.10: Contribution of each term in Eq. (6.9) to the full variance, for multiple wavelength bins. For this measurement we only used the 5% of quasars with the highest SNR in the EDR+M2 sample. This analysis reveals that the effect of the  $\epsilon$  term in Eq. (6.9) is minimal, even for the data subset where it is supposed to be most significant.

Nevertheless, for this DESI EDR analysis, we decided to omit this modification since its inclusion lacked sufficient justification, and its influence on the analysis is negligible.

## 6.4 Discussion

The main science driver of the Lyman- $\alpha$  catalog presented here is the measurement of 3D correlations that allows us to constraint the BAO scale (Gordon et al., 2023). In this Section, we discuss two methodological novelties in the construction of the Lyman- $\alpha$  catalog with respect to previous eBOSS analyses (du Mas des Bourboux et al., 2020), and we do this by looking at the impact of these changes to the precision with which we will be able to measure the 3D correlations. In particular, in Section 6.4.1 we discuss the optimization of the weights assigned to each Lyman- $\alpha$  fluctuation, while in Section 6.4.2 we will choose the rest-frame wavelength range based on the precision of the correlation function measurements.

We conclude this Section in Section 6.4.3 with a discussion on the distribution of per-quasar parameters that capture the diversity of quasar continua in the dataset.

### 6.4.1 Optimal weights

Correlations in the Lyman- $\alpha$  forest are estimated as weighted averages of products of fluctuations  $\delta_q(\lambda)$  in pixel pairs at a given separation. An optimal quadratic estimator would use the inverse of the pixel covariance as the weight matrix, but the large number of pixels in current Lyman- $\alpha$  datasets makes this inversion not feasible.

Ignoring the small correlation between pixels in different quasar spectra, one can approximate the covariance as block-diagonal, and make the inversion tractable. This approximation has been used in several measurements of the 1D power spectrum (McDonald et al., 2006; Karaçaylı et al., 2020, 2022, 2024), it has been proposed to measure the 3D power spectrum (Font-Ribera et al., 2018) and it was used in one of the first BAO measurements with the Lyman- $\alpha$  forest (Slosar et al., 2013).

Recent measurements of 3D correlations in the Lyman- $\alpha$  forest (Delubac et al., 2015; Bautista et al., 2017; de Sainte Agathe et al., 2019; du Mas des Bourboux et al., 2020), on the other hand, have ignored all correlations between pixels and have used instead a diagonal weight matrix. These studies weighted each pixel with the inverse of its variance, including the instrumental noise and the intrinsic fluctuations (see Eq. (6.6)), effectively approximating the inverse covariance matrix with the inverse of its diagonal elements. This approximation is simple and easy to implement, but it is not the optimal diagonal weight matrix.

In this Section we study a simple modification of our diagonal weight matrix, where we add an extra free parameter  $\sigma_{\text{mod}}^2$  that modulates the contribution of the intrinsic fluctuations  $\sigma_{\text{LSS}}$  to the weights:

$$w_q(\lambda) = \frac{1}{\eta(\lambda)\tilde{\sigma}_{\text{pip},q}^2(\lambda) + \sigma_{\text{mod}}^2\sigma_{\text{LSS}}^2(\lambda)} . \quad (6.10)$$

A small value of  $\sigma_{\text{mod}}^2$  is equivalent to weighting the pixels based solely on the noise variance, while a large value gives the same weight to all pixels at a given redshift, regardless of instrumental noise.

In Fig. 6.11 we show that a value of  $\sigma_{\text{mod}}^2$  around 7-8 can improve the precision of the auto-correlation measurement by 25%, at no additional cost. As expected, the gain in precision in the cross-correlation is roughly half of that, and we find a 10% improvement for values of  $\sigma_{\text{mod}}^2$  around 6-7.

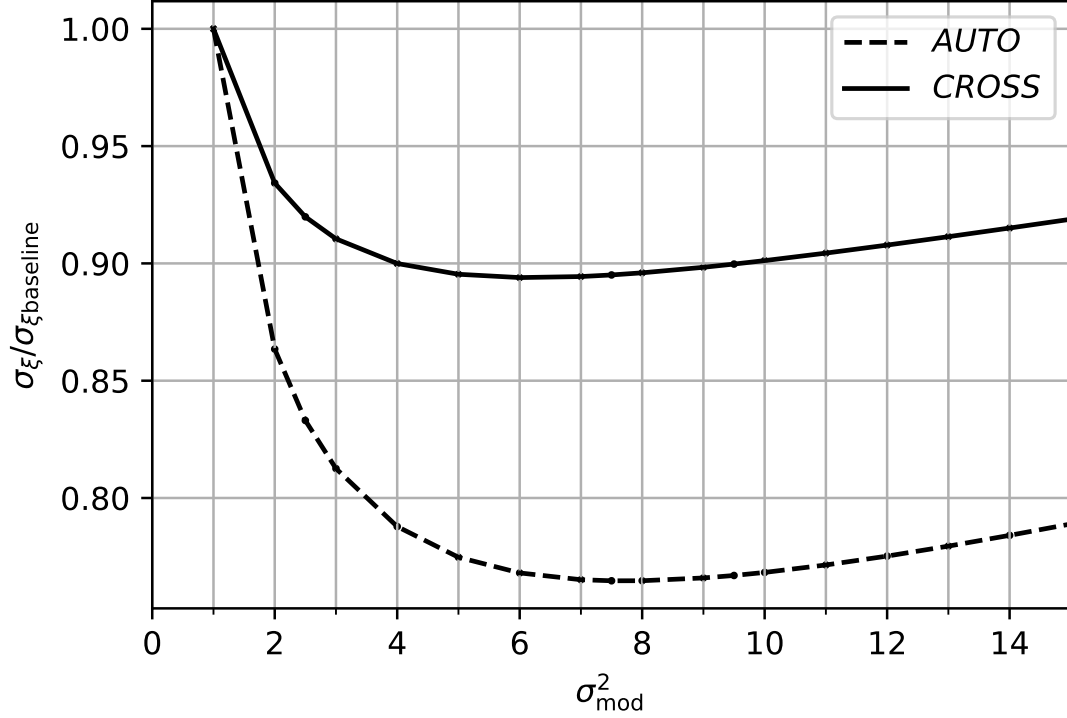


Figure 6.11: Measurement of the Lyman- $\alpha$  auto- (solid) and cross- (dashed) correlation errorbars for different choices of the  $\sigma_{\text{mod}}^2$  parameter, scaled to a reference value at  $\sigma_{\text{mod}}^2 = 1$ . The value of the errorbars was averaged over all scales, given the small scale dependency. The  $\sigma_{\text{mod}}^2$  parameter modified the inverse-variance weighting scheme as defined in Eq. (6.10), the use of a  $\sigma_{\text{mod}}^2 \neq 1$  allows us for the optimization of the weighting scheme. In both auto- and cross-correlations, we observe a reduction in the size of the errorbars when we approach the optimal value of the  $\sigma_{\text{mod}}^2$  parameter. This optimal value is slightly different, but in both cases is found around 7-8. In the optimal value, the improvement in the measurement of the auto-correlation is about 20%, and 10% for the case of the cross-correlation. Both measurements have been performed with the full EDR+M2 dataset.

It is important to note that the actual gain will depend on the properties of the dataset. For instance,  $\sigma_{\text{pip}}$  and  $\sigma_{\text{LSS}}$  change differently with the width of the pixels used, and we have tested that when using pixels of  $2.4\text{\AA}$  (similar to the ones used in the BAO measurement of du Mas des Bourboux et al. (2020)) the optimal value is smaller,  $\sigma_{\text{mod}}^2 = 3.1$ , and the gain in the auto-correlation is only 8%.

For this work and the associated Value Added Catalog, we decide to use a value of  $\sigma_{\text{mod}}^2 = 7.5$ , and we leave for future work the implementation of a block-diagonal weighting.

### 6.4.2 Selection of rest-frame wavelength range

Due to diversity in quasar spectra near the emission lines, the rest-frame wavelength range that can be used for analyzing the Lyman- $\alpha$  forest is limited. This happens at the red end of the spectra due to the Lyman- $\alpha$  emission line ( $\lambda_{\text{RF}} = 1215.67\text{\AA}$ ) and at the blue end limited due to the Lyman- $\beta$  line ( $\lambda_{\text{RF}} = 1025.72\text{\AA}$ ).

Extending the analysis towards longer wavelengths closer to the Lyman- $\alpha$  emission line allows for the incorporation of more data. By including these extra pixels in our analysis, we can improve our measurements of the correlation function if the improvement due to the large number of pixels is not counterbalanced by of the larger pixel variance.

In Fig. 6.12, the blue points show the size of the errorbars in the auto-correlation measurement when we extend the analysis to higher  $\lambda_{\text{RF, max}}$ . The value at  $\lambda_{\text{RF, max}} = 1205\text{\AA}$  yields the smallest errorbars. To separate the two effects of increased number of pixels and increased pixel variance, we also plot  $\sigma_{\xi}^2 N$ , where  $N$  is the number of pairs in the correlation measurement, since we expect  $\sigma^2 \propto 1/N$ . This will show how valuable the added points are when extending the wavelength range. Its evolution in Fig. 6.12 shows that extending  $\lambda_{\text{RF, max}}$  to higher wavelengths adds less valuable information, and therefore the decrease in the size of the errorbars is only driven by the increase in the sample size.

Given this result, we decided to set  $\lambda_{\text{RF}} = 1205\text{\AA}$  because we found that further increasing the limit did not add constraining power. The increased variance near the emission line is likely the reason behind this, as it is not accounted for in our calculations and could potentially affect our measurements if we approached it too closely. A more detailed study of quasar continuum variance, using a larger dataset than the one presented here, is necessary to fully understand its effects. We leave this for future releases of DESI data.

The same exercise was performed for the blue side of the quasar spectrum, and the results are shown in Fig. 6.13. When compared to the prescription of previous

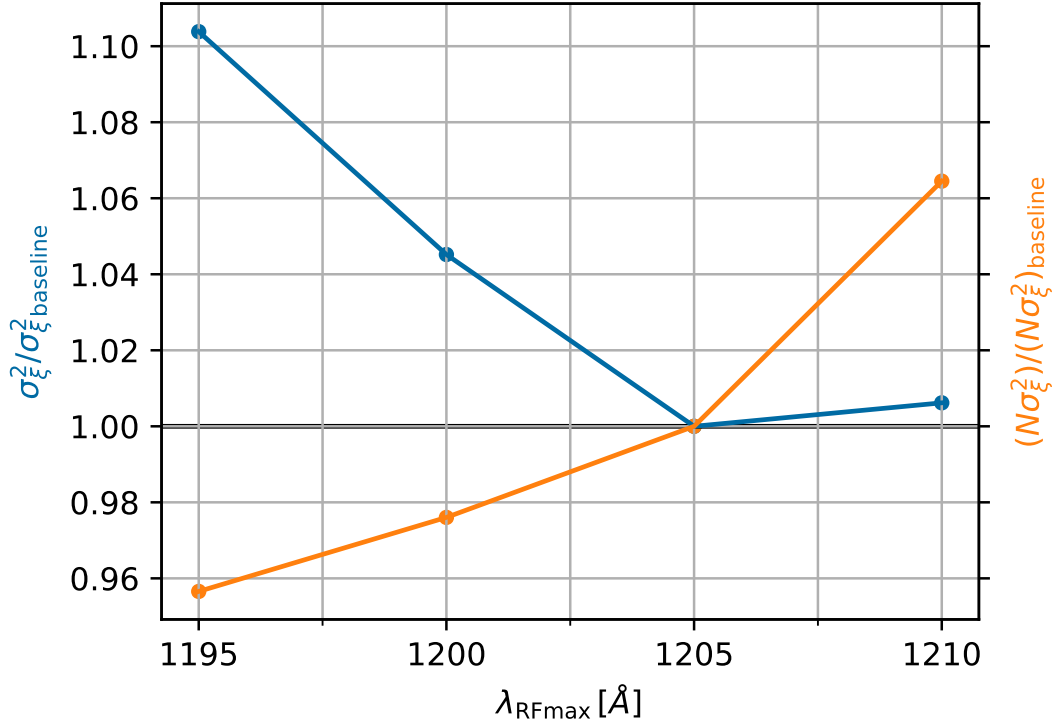


Figure 6.12: Comparison of the size of the errorbars in the Lyman- $\alpha$  auto-correlation for different choices of  $\lambda_{\text{RF,max}}$ . The blue points show the errorbars size after averaging over all scales (given the small scale dependence of this quantity), and scaled to a reference value at  $\lambda_{\text{RF,max}} = 1205 \text{ \AA}$ . The orange points show the equivalent measurement but in this case for the product of errorbar sizes times the number of pixel pairs, removing the dependence on the number of pairs used in the measurement. Following the blue points, we observe an improvement in the precision of our auto-correlation measurements when increasing the  $\lambda_{\text{RF,max}}$  parameter, having the optimal value at  $1205 \text{ \AA}$ . The orange points show that the quality of these added points actually decrease for higher wavelengths, revealing that the improved performance is only driven by the inclusion of more information in our sample. We selected  $1205 \text{ \AA}$  as our default value as a compromise between these two features. We used data from the whole EDR+M2 dataset for these measurements.

analyses ( $\lambda_{\text{RF,min}} = 1040 \text{ \AA}$  in du Mas des Bourboux et al. (2020)), we observe a degradation of the errorbars either going to smaller  $\lambda_{\text{RF,min}}$  due to the decrease of pixel count; or to larger  $\lambda_{\text{RF,min}}$  approaching the emission line. For this reason we retain this prescription of  $\lambda_{\text{RF,min}} = 1040 \text{ \AA}$ .

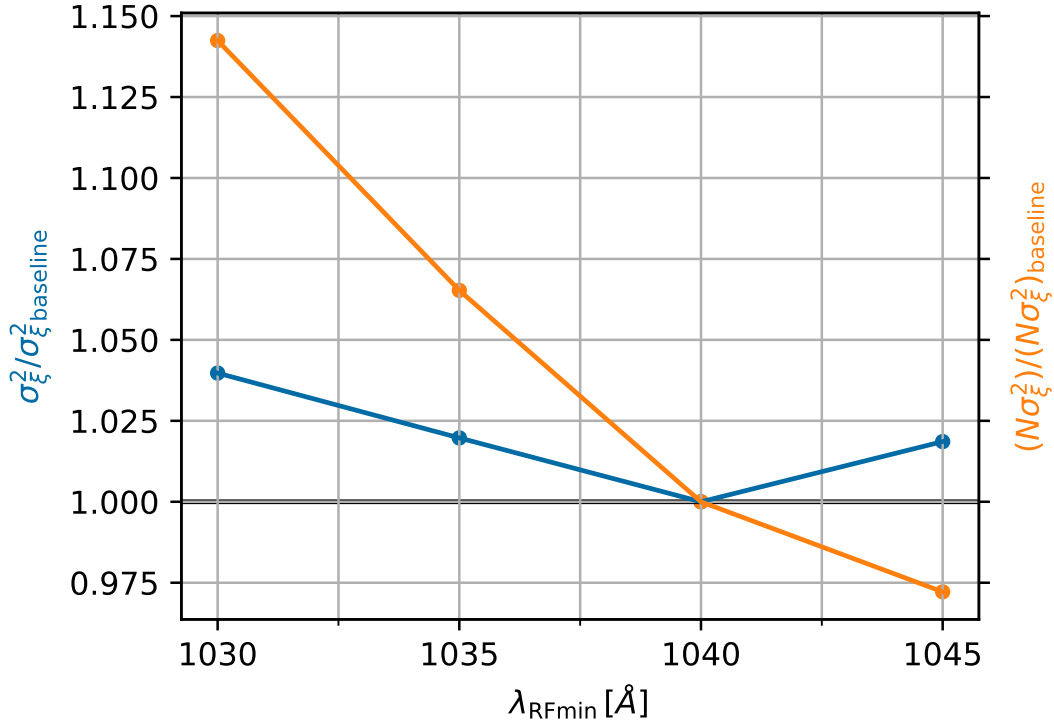


Figure 6.13: Identical measurements as the ones displayed in Fig. 6.12, in this case for values of  $\lambda_{\text{RF, min}}$ . The optimal wavelength is found at  $\lambda_{\text{RF, min}} = 1040 \text{ \AA}$ , while having similar features for the orange points: a decrease in the quality of the points added when approaching the emission line. Following the same arguments as in the case of  $\lambda_{\text{RF, max}}$ , we decided to keep the limit at  $1040 \text{ \AA}$ . Again, these measurements were performed using the EDR+M2 sample.

### 6.4.3 Per-quasar parameters (a,b)

As mentioned in Section 6.3.3, the quasar continua is assumed to follow a universal function of rest-frame wavelength, with a correction by a first degree polynomial parametrized by  $a_q$  and  $b_q$ . Quasar variability can be larger at both ends of the Lyman- $\alpha$  forest, due to the Lyman- $\beta$  and Lyman- $\alpha$  emission lines, but variability in the weak quasar emission lines at  $1017 \text{ \AA}$  and  $1123 \text{ \AA}$  (Suzuki, 2006) could also impact the results.

In Fig. 6.14, we examine the distribution of these two parameters to test this assumption. The plot does not show special features apart from the long tail at large values of  $|b_q/a_q|$ . This feature is caused by spectra where only a small part of the forest appears in the spectrograph. In these cases, fitting the real continuum of the quasar is difficult, leading to poor fits. For this reason, we require forests to have a minimum length of 150 pixels of  $0.8 \text{ pixels}$ .

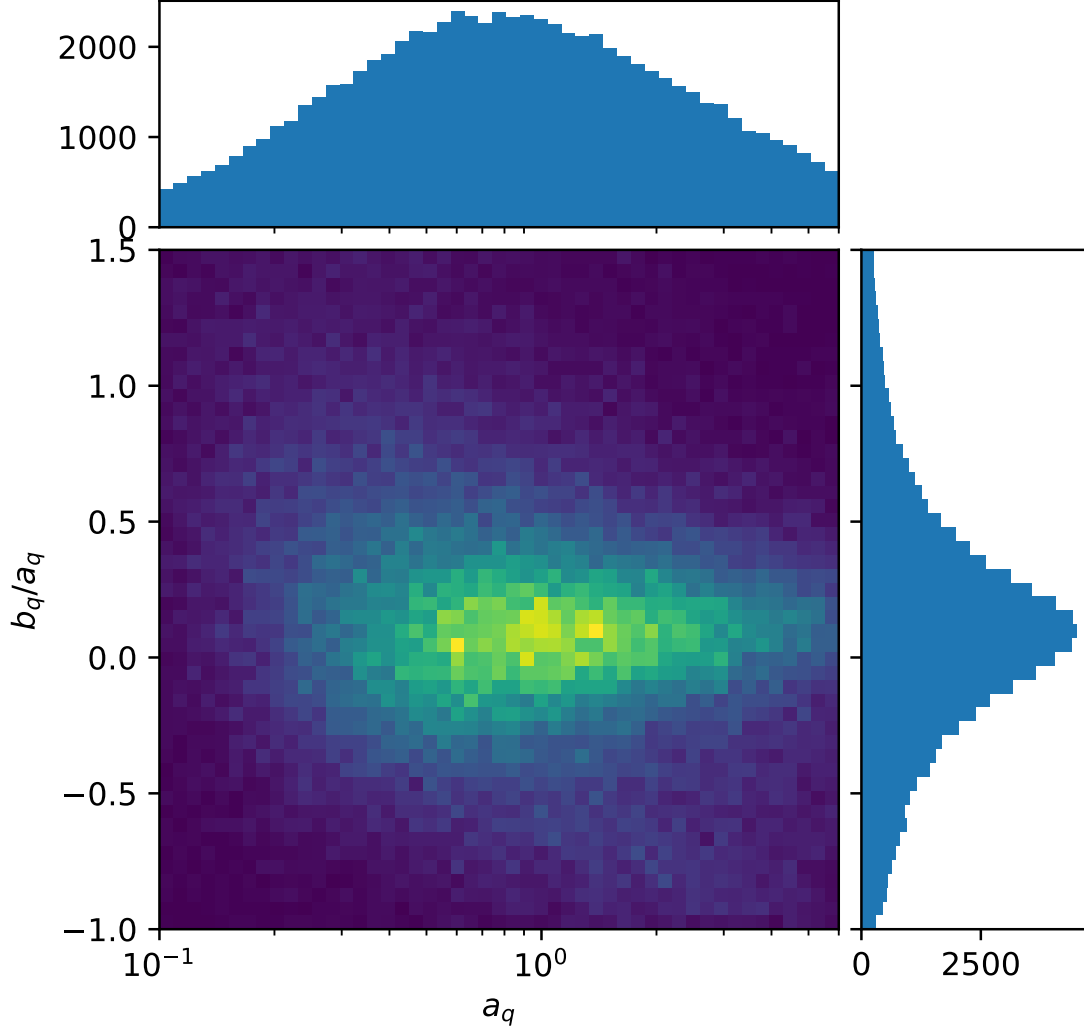


Figure 6.14: Distribution of  $a_q$  and  $b_q/a_q$  parameters defined in Eq. (6.4) for the EDR+M2 sample. The  $a_q$  parameter modifies the amplitude of the mean continuum  $\overline{C}(\lambda_{\text{RF}})$ , while the  $b_q/a_q$  term introduces a modification that tilts it as a function of rest-frame wavelength. This last parameter relates to the intrinsic width of the spectral index. The faint long tail at large values of  $|b_q/a_q|$  is caused by short forests in the sample, where the continuum fit can be problematic.

## 6.5 Summary and conclusions

In this work, we present the first measurement of Lyman- $\alpha$  fluctuations using DESI data. We use two data samples: EDR, which mainly consists of SV data; and EDR+M2, which also includes the first two months of the main survey (M2) data. The EDR sample contains 68 750 quasars, of which 20 281 have valid forests for Lyman- $\alpha$  studies, i.e., with the Lyman- $\alpha$  region visible and identifiable. The EDR+M2 sample contains 318 691 quasars, of which 88 511 have valid forests. We release the Lyman- $\alpha$  fluctuations catalog for the EDR+M2 sample as part of the first DESI data release.

To achieve this measurement, we adapted the methodology proposed in previous analyses, especially the one outlined in du Mas des Bourboux et al. (2020), to suit the unique characteristics of the DESI early data sample. We discussed and applied several important modifications.

- We adjusted our pipeline to work with linear-spaced bins in wavelength to match the DESI scheme, preserving its format and precision.
- We adapted the re-calibration process by simplifying it, removing unneeded steps that did not improve the performance of the whole re-calibration. We also switched the re-calibration region to allow for more pixels to be used in this process.
- We simplified the weighting scheme by removing terms that added unnecessary complexity and did not contribute to the accuracy of the results, specially for this small data release.
- By optimizing our diagonal weight matrix, we improved measurements of the auto-correlation by about 20%, and of the cross-correlation by about 10%

The presented flux-transmission field catalog could be used for other studies apart from the standard BAO analysis. As in Font-Ribera et al. (2012b) and Pérez-Ràfols et al. (2018a,b), its cross-correlation with DLAs can be measured. A similar analysis could be performed with Strong Blended Lyman- $\alpha$  (SBLA) absorption systems, as in Pérez-Ràfols et al. (2023). The full-shape analysis of the Lyman- $\alpha$  forest three-dimensional correlation function would allow for the measurement of the Alcock-Paczyński effect, as performed in Cuceu et al. (2023a) using eBOSS DR16 data. Furthermore, IGM tomography is also possible with this sample.

As datasets get larger and larger, we will need to be more careful with the analysis. Weights used to measure clustering could be improved in two ways.



As discussed in Section 6.4.1, the implementation of a block-diagonal weighting scheme will account for correlations between pixels within the same forest. Alternatively, taking into account that quasar diversity makes the estimation of the continua more difficult, one possible way of improving the weights is adding to the weighting scheme proposed in Eq. (6.6) an extra term accounting for errors in the estimation in the continuum. This term would be dependent on  $\lambda_{\text{RF}}$  and would be larger around the emission lines, where quasar diversity is expected to be higher. Finally, the analysis could be extended into the Lyman- $\beta$  region.

This study can serve as a solid foundation for future research using DESI data. We are excited about the potential for further investigations in this area, as more comprehensive analyses become possible with the availability of future DESI releases.

# ROBUSTNESS TESTS FOR THE DESI DR1 LYMAN- $\alpha$ FOREST BAO ANALYSIS

In Chapter 6, we showed how the Lyman- $\alpha$  catalog for the Early Data Release (EDR) was designed and built. In April 2024, the Dark Energy Spectroscopic Instrument (DESI) collaboration released the Baryon Acoustic Oscillations (BAO) results from the Data Release 1 (DR1), containing data from the first year of main survey. This included measurements from galaxies and quasars (DESI Collaboration et al., 2024a) and from the Lyman- $\alpha$  forest (DESI Collaboration et al., 2024b). Cosmological constraints using these measurements were released in DESI Collaboration et al. (2024c).

The Lyman- $\alpha$  dataset contained information from more than 700 000 quasars, twice the number of previous BOSS+eBOSS analyses (du Mas des Bourboux et al., 2020). This release provided a measure of the expansion at  $z_{\text{eff}} = 2.33$  with 2% precision and a 2.4% measurement of the transverse comoving distance. Similar methods as the ones in the EDR were used, specifically, the Lyman- $\alpha$  forest catalog was generated using the procedures detailed in Chapter 6.

In this Chapter, we will describe part of the validation necessary to perform the DR1 analysis. Specifically, the tests conducted to ensure the robustness of the analysis, consisting of variations around the main analysis and some data splits. The goal was to ensure that the measurement of the BAO position did not depend on specific configurations of the analysis. These analyses were published alongside the Lyman- $\alpha$  forest measurements in (DESI Collaboration et al., 2024b).

BAO measurements are particularly robust because they are based on the measurement of a well-defined three-dimensional feature, while the spurious correlations caused by instrumental systematics or contaminants tend to be smooth

and featureless<sup>1</sup>.

Despite this, the complexity of the analysis and its multiple contaminants mean that the BAO measurement could be affected by decisions made during the analysis setting, such as the selection of the wavelength range, or the masking discussed in Section 6.3.1. Therefore, we must ensure that all these decisions do not impact the final results.

The DESI Collaboration requires all the working groups working on BAO analyses to work with blinded data, aiming to avoid confirmation biases. The blinding strategy chosen for Lyman- $\alpha$  forest analyses was one where the BAO peak was shifted by an unknown additive term (for more details on the blinding strategy, see Appendix C of DESI Collaboration et al. (2024b)).

The blinding strategy included conditions that must be met before unblinding. Ideally, after the unblind, no further modifications should be made without heavily justifying them beforehand. In our case, the unblinding condition was that a series of tests conducted on both mocks and real data did not yield results significantly different from the main analysis. The relevant statistic chosen for this was the position of the BAO peak using the scale parameters  $\alpha_{\parallel}$  and  $\alpha_{\perp}$ , defining the position in the direction parallel and perpendicular to the line of sight respectively, defined as:

$$\alpha_{\parallel} = \frac{D_H(z)/r_d}{(D_H(z)/r_d)_{\text{fid}}}, \quad \alpha_{\perp} = \frac{D_M(z)/r_d}{(D_M(z)/r_d)_{\text{fid}}}, \quad (7.1)$$

where the fid subscript indicates the values for the fiducial cosmology used in the analysis to compute comoving separations from angles and redshifts.

The validation in synthetic data is extensively discussed in the companion paper focused on mock data for DR1. This synthetic datasets included both datasets coming from CoLoRe realisations (software that we presented in Chapter 5) and an independent software developed independently and presented in Etourneau et al. (2024), the Saclay mocks.

The validation with real data for BAO measurements can be divided into two parts: data splits, where the data are divided into qualitatively different parts; and alternative analyses, where analyses are conducted with different configurations. In both cases, we seek to ensure that the BAO results are consistent.

---

<sup>1</sup>One exception is the case of contamination by absorption by elements other than H. The correlation between these absorptions will cause peaks in the correlation function that need to be properly modelled (see Gordon et al. (2023) for details)

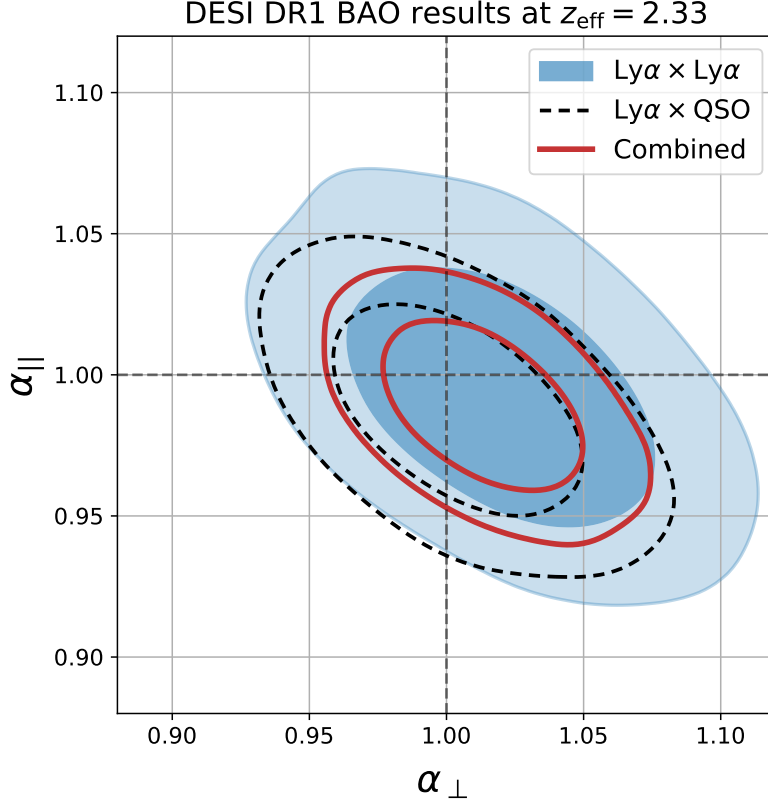


Figure 7.1: Measurements of the BAO parameters along the line of sight ( $\alpha_{\parallel}$ ) and across the line of sight ( $\alpha_{\perp}$ ) with contours corresponding to the 68% and 95% confidence regions. The auto-correlation results (filled blue contours) are the combined measurement of the Ly $\alpha$  forest auto-correlations in the Lyman- $\alpha$  and Lyman- $\beta$  regions. The cross-correlation results (dashed black) are the correlations of the forest in these two regions with quasars. The combined results (solid red) simultaneously fit all four correlations taking into account their cross-covariance (DESI Collaboration et al., 2024b).

## 7.1 Data splits

In the data splits, we divide the real data into two qualitatively different parts and measure the impact on BAO by comparing it with the main analysis. It is important to note that when performing a data split, one ends up with two smaller subsets of data, and we should expect a larger statistical uncertainty.

The first data split performed corresponds to the results shown in Fig. 7.1, where we compare the consistency of BAO measurements by comparing the Lyman- $\alpha$  auto-correlation and the cross-correlation with quasars. Fig. 7.1 shows the contours (68% and 95% confidence regions) in the  $(\alpha_{\parallel}, \alpha_{\perp})$  space for the auto-correlation, the cross-correlation, and the combined result.

The rest of the data splits can be seen in Fig. 7.2. In the top left panel, we see a split based on the mean Signal-to-noise ratio (SNR) of the spectrum. The chosen

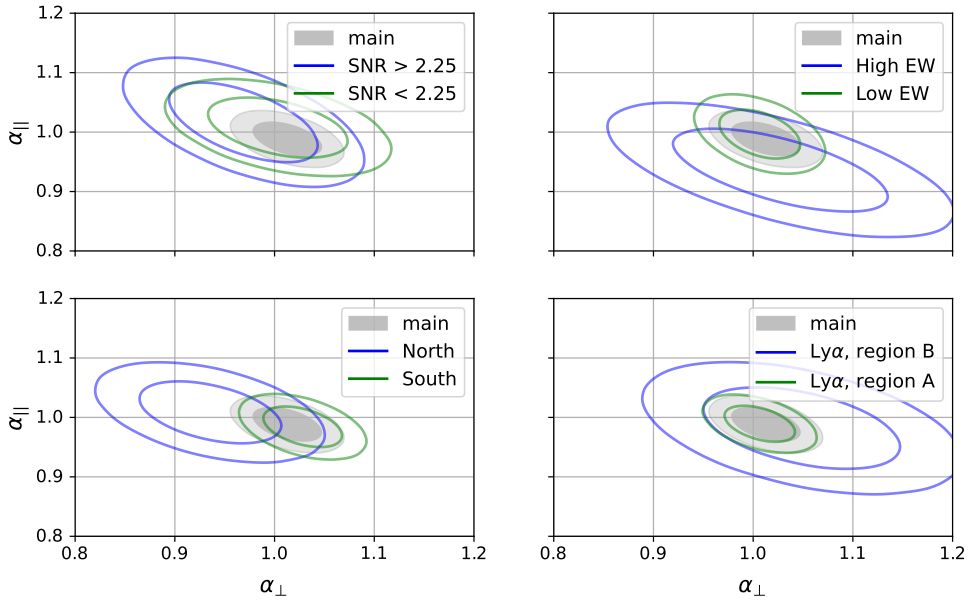


Figure 7.2: BAO constraints from the main analysis (grey) and from data splits. Top left: low (green) vs high (blue) SNR in the quasar spectrum. Top right: low (green) vs high (blue) C IV equivalent width (EW) in the quasar spectrum. Bottom left: South (green) vs North (blue) imaging used in the quasar target selection. Bottom right: correlations from the Lyman- $\alpha$  region (green) and Lyman- $\beta$  region (blue); the Lyman- $\alpha$  region shows the combined measurement from the auto-correlations of the forest measured in the Lyman- $\alpha$  region and the cross-correlations of this region with quasars. The contours labeled Lyman- $\beta$  show the combined measurement of the forest auto-correlations measured in the Lyman- $\beta$  region and the cross-correlation of this region with quasars.

value of 2.25 is selected so that the two parts of the split have roughly the same weight in the auto-correlation measurement. The SNR value is the average in the Lyman- $\alpha$  region. The two resulting catalogs have 321 767 quasars for the low SNR split and 106 636 for the high SNR split. The sum of both values does not reach the total number of quasars in the general sample (709 565) because it is not possible to detect the forest continuum for  $z < 2$  and therefore to obtain an estimate for SNR.

In the top right panel of the same Fig. 7.2, we show a split where we divide the quasar sample based on the width of the C IV emission line (EW). This division is motivated by the expectation of a dependence between the EW and the continuum luminosity, a phenomenon known as the Broad Absorption Line (BAL)dwain Effect (Baldwin, 1977). The width measurement gives an average value of 37.3 Å, with a mean of 41.6 Å for  $3\sigma$  measurements of the emission line. Based on this data,

the sample is split at  $39 \text{ \AA}$ , resulting in 371 751 quasars for the low EW sample and 337 814 for the high EW sample. It has also been observed that the low EW sample has a slightly higher luminosity than the high EW sample (as predicted by the BALdwin Effect).

The next split, shown in the bottom left panel of Fig. 7.2, shows the results for a split in the footprint. This split is based on differences in the imaging survey used for target selection. A small part of the footprint is based on observations from the BASS and MzLS surveys (Silva et al., 2016; Zou et al., 2017) in the North Galactic Cap, for  $\delta > 32.375^\circ$  (where  $\delta$  the declination), we have considered this part as the “North” split. On the other hand, most of the DESI footprint comes from target selection performed by the DECam camera on the Blanco telescope in Chile, including the entire South Galactic Cap and the southern part of the North Galactic Cap. We have considered this other part of the split as “South”. The “South” sample is considerably larger than the north, containing 579,166 quasars compared to 130,399 in the north split.

Finally, shown in the bottom right of the same Fig. 7.2, we have the split considering different regions of the Lyman- $\alpha$  forest. In the main Lyman- $\alpha$  forest analysis, we consider two regions to calculate the correlations of the fluctuations. The first region, called Lyman- $\alpha$ , comprises wavelengths between  $1040 \text{ \AA}$  and  $1215 \text{ \AA}$  in the quasar rest-frame wavelength, immediately to the left of the Lyman- $\alpha$  emission line. This region obtains the “purest” measurements as there is no contamination from other hydrogen emission lines. On the other hand, we have the Lyman- $\beta$  region, which comprises wavelengths between  $920 \text{ \AA}$  and  $1020 \text{ \AA}$ , this is to the left of the Lyman- $\beta$  emission line.

In this data split, we use Lyman- $\alpha$  absorption in the Lyman- $\alpha$  region and Lyman- $\alpha$  absorption in the Lyman- $\beta$  region. Because the wavelengths in the Lyman- $\beta$  region are shorter, this region only appears in higher redshift quasars. This implies that we have fewer data for this second region, contributing to a larger variance in the result of the measurement in the Lyman- $\beta$  region.

With the exception of the splits separating the North from the South, the others share the same footprint and redshift range. Despite this, the cosmic variance is a very small contribution to the covariance matrix, so we can approximate the splits as independent. Observing the contours in Fig. 7.1, Fig. 7.2, we can conclude that the different tests are compatible with statistical fluctuations.

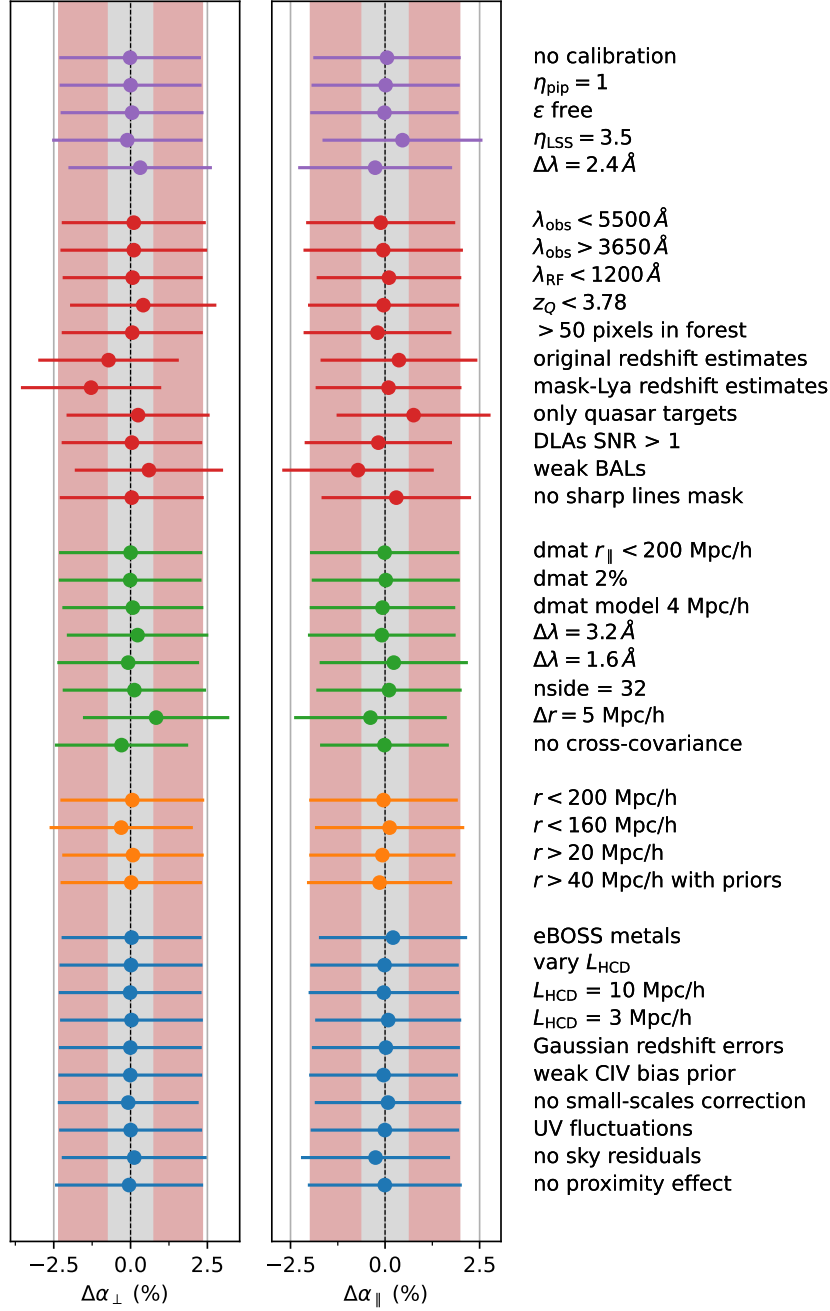


Figure 7.3: Shifts in the BAO parameters from alternative analyses. These include variations in the method to estimate the fluctuations (purple); variations in the dataset used (red); variations in the measurement of correlations and covariances (green); variations in the range of separations used (orange); and variations in the modelling (blue). The red shaded region shows the one  $\sigma$  uncertainty from the main analysis, and the smaller gray region shows the threshold set for these tests ( $\sigma/3$ ).

## 7.2 Alternative analyses

Now we will discuss the validation tests focused on the robustness of the BAO measurement. In this type of test, we modify some configurations of the main pipeline to check if there is any shift in the BAO parameters. Before unblinding, we set a value of one-third of the uncertainty in the main analysis as a threshold to test the robustness. This corresponds in absolute terms to  $\alpha_{\parallel} \sim 0.005$  and  $\alpha_{\perp} \sim 0.007$ .

However, some of the variations used affect the sample size. In this type of variation, we expect a larger statistical uncertainty, and it is possible that some of the tests may fail to meet the parameters stipulated in the previous paragraph.

There are three types of variations depending on the step in the pipeline where the change is introduced: variations in the estimation of the fluctuations, variations in the measurement of the correlations, and variations in the estimation of cosmological and astrophysical parameters.

### 7.2.1 Variations in the estimation of fluctuations

These variations affect the estimation of fluctuations from the quasar spectra provided by the DESI pipeline.

First we have variations where the methodology used in the estimation of fluctuations is changed, these are shown in purple in Fig. 7.3.

- no calibration: In this variation, we do not perform the calibration step in the C III region as described in Section 6.3.2.
- $\eta_{\text{pip}} = 1$ : We do not add the correction to the pipeline noise, as described in Section 6.3.3.
- $\epsilon$  free: We recover the  $\epsilon$  term described in Section 6.3.4.2, used in previous Extended Baryon Oscillation Spectroscopic Survey (eBOSS) analyses to capture quasar diversity.
- $\sigma_{\text{mod}}^2 = 3.5$ : We reduce the contribution of the intrinsic Lyman- $\alpha$  forest variance to the weights of each fluctuation value, as defined in Section 6.4.1.
- $\Delta\lambda = 2.4 \text{ \AA}$ : We coadd the spectra in bins 3 times larger than the original, as was done in previous eBOSS analyses. In this case, we use  $\sigma_{\text{mod}}^2 = 3.1$ , the suggested value in Section 6.4.1.

We also have variations where we choose different cuts in the quasar spectra, shown in Fig. 7.3 in red. In this type of variation, the dataset size is smaller, and



therefore we expect slightly larger statistical fluctuations compared to the cases described earlier:

- $\lambda_{\text{obs}} < 5500 \text{ \AA}$ : We use only wavelength pixels below this value in observed wavelength (default is  $5577 \text{ \AA}$ ).
- $\lambda_{\text{obs}} > 3650 \text{ \AA}$ : We use only wavelength pixels above this value in observed wavelength (default is  $3600 \text{ \AA}$ ).
- $\lambda_{\text{RF}} < 1200 \text{ \AA}$ : We use only wavelength pixels below this value in rest-frame wavelength (default is  $1205 \text{ \AA}$ ).
- $z_{\text{QSO}} < 3.78$ : We use only quasars with redshift values up to this maximum (matching the maximum value available in the mocks).<sup>2</sup>
- $> 50$  pixels in forest: We only use quasars whose spectrum has more than 50 valid pixels in the interest regions, compared to the value of 150 in the main analysis.
- original redshift estimates: We use the original redshift values of the quasars provided by Redrock. It was shown in Brodzeller et al. (2023) that these values are slightly biased.
- mask-Lyman- $\alpha$  redshift estimates: Similar to the eBOSS analysis, we mask regions below the Lyman- $\alpha$  emission line when determining the quasar redshift.
- only quasar targets: We use a quasar catalog where only objects targeted as quasars are included.
- Damped Lyman- $\alpha$  Absorption (DLA)s  $\text{SNR} > 1$ : In the main analysis, we mask and correct for DLAs found in spectra with  $\text{SNR} > 3$ . For this variation, we lower the threshold to 1. The masking process was described in Section 6.3.1.2.
- weak BALs: We exclude quasars from the analysis that exhibit strong BAL features ( $A_{\text{I}} > 840$ , 50th percentile of strongest BALs in eBOSS Ennesser et al. (2022)). BAL masking was described in Section 6.3.1.2.
- no sharp lines mask: We do not mask the sharp lines defined in Section 6.3.1.

---

<sup>2</sup>A small difference from the EDR analysis is that in DR1 there was no redshift limit imposed on quasars.

### 7.2.2 Variations in the measurement of correlations

In the previous Section we saw variations involving features mostly discussed in Chapter 6. In the next two Sections, we will see variations involving the measurement of correlations and in the estimation of cosmological and astrophysical parameters. These two types of variations involve features not heavily discussed in this Thesis, we point to Gordon et al. (2023) for a better understanding of the correlation measurement process and to DESI Collaboration et al. (2024b) for the estimation of the cosmological and astrophysical parameters.

The variations where the configuration used to measure correlations is modified are shown in Fig. 7.3 in green color and are the following ones:

- $\text{dmat } r_{\parallel} < 200 \text{ Mpc/h}$ : We model the distortion matrix only up to  $r_{\parallel} = 200 \text{ Mpc/h}$ , compared to the default value of  $300 \text{ Mpc/h}$ . This lower value was used in the eBOSS analysis.<sup>3</sup>
- $\text{dmat } 2\%$ : We use 2% of the pixels to calculate the distortion matrix (1% in the main analysis).
- $\text{dmat model } 4 \text{ Mpc/h}$ : We model the distortion matrix using the same binning as the correlation function (using the finer binning of  $2 \text{ Mpc/h}$  in the main analysis).
- $\Delta\lambda = 3.2 \text{ \AA}$ : We rebin the fluctuations in groups of 4 pixels before calculating correlations (3 in the main analysis).
- $\Delta\lambda = 1.6 \text{ \AA}$ : We rebin the fluctuations in groups of 2 pixels before calculating correlations (3 in the main analysis).
- $\text{nside} = 32$ : The correlation function is calculated in different Healpix pixels to compute covariances. In this case, we generate pixels with  $\text{nside}=32$  instead of the main analysis's  $\text{nside}=16$ .
- $\Delta r = 5 \text{ Mpc/h}$ : We use a thicker bin for the correlation function compared to the  $4 \text{ Mpc/h}$  bin used in the main analysis.
- $\text{no cross-covariance}$ : There exists a covariance between auto-correlations and cross-correlations used<sup>4</sup>. In this variation, we do not include it in the analysis.

<sup>3</sup>The distortion matrix accounts for the distortion that occurs in measurements of Lyman- $\alpha$  fluctuations when estimating the quasar continuum. See Gordon et al. (2023) for more details.

<sup>4</sup>See DESI Collaboration et al. (2024b) or Cuceu et al. (2024) for more details on this cross-covariance.

In all these variations, there is no significant difference compared to the main analysis in the position of the BAO parameter.

### 7.2.3 Variations in the estimation of cosmological and astrophysical parameters

In this Section, we address the impact of variations in parameter estimation. On the one side, we have variations where the range of separations is changed, corresponding to the orange color in Fig. 7.3:

- $r < 200$  Mpc/h: We fit separations only between 10 and 200 Mpc/h (compared to 180 Mpc/h in the main analysis).
- $r < 160$  Mpc/h: We fit separations only between 10 and 160 Mpc/h (compared to 180 Mpc/h in the main analysis).
- $r > 20$  Mpc/h: We fit separations between 20 and 180 Mpc/h (compared to 10 Mpc/h in the main analysis).
- $r > 40$  Mpc/h with priors: We fit separations between 40 and 180 Mpc/h (compared to 10 Mpc/h in the main analysis). Since we are cutting off smaller scales, it is challenging to constrain several nuisance parameters. To alleviate this, we use priors on the more difficult to estimate parameters<sup>5</sup>.

On the other hand we have variations where different decisions have been made regarding the model, shown in blue color in Fig. 7.3:

- eBOSS metals: Heavier elements absorptions need to be properly modelled to account for their effects on the correlation function. In this variation, we use the eBOSS method instead of the new method described in DESI Collaboration et al. (2024b).
- vary  $L_{\text{HCD}}$ : In this variation, we leave the parameter  $L_{\text{HCD}}$  free in the analysis, which was fixed at 6.51 Mpc/h in the main analysis. Even though DLAs are masked in the analysis, there are some systems that we miss, and we can model their effect (du Mas des Bourboux et al., 2020). This parameter controls the characteristic width of DLA systems.
- $L_{\text{HCD}} = 10$  Mpc/h: We fix this parameter to a higher value than in the main analysis (6.51 Mpc/h).

---

<sup>5</sup>See DESI Collaboration et al. (2024b) for details on the parameters where priors were applied

- $L_{\text{HCD}} = 3 \text{ Mpc/h}$ : We fix this parameter to a smaller value than in the main analysis (6.51 Mpc/h).
- Gaussian redshift errors: The errors in the estimation of quasar redshift are assumed to follow a Gaussian distribution. In the main analysis, a Lorentzian distribution is used.
- weak C IV bias prior: We use a more relaxed prior for the C IV bias parameter<sup>6</sup>, using  $-0.03 < b_{\text{C IV}} < 0$  instead of  $-0.0243 \pm 0.0015$  in the main analysis.
- no small-scales correction: We ignore the small-scales multiplicative corrections from Arinyo-i-Prats et al. (2015) in the Lyman- $\alpha$  auto-correlation.
- UV fluctuations: Following the analysis in Bautista et al. (2017), we model the impact of fluctuations in the UV background on the Lyman- $\alpha$  forest auto-correlation. We did not detect this effect in our analysis.
- no sky residuals: We ignore the contamination from correlated sky residuals in the Lyman- $\alpha$  auto-correlation, effect discussed in DESI Collaboration et al. (2024b).
- no proximity effect: We ignore the impact of quasar radiation in the cross-correlation, effect discussed in DESI Collaboration et al. (2024b).

Finally, we performed one last variation where we included broadband polynomial corrections. These corrections involve adding a smooth additive component to all correlations. We used Legendre polynomials  $L_j(\mu)$  up to order  $j = 0, 2, 4$ , and 6 to describe the angular dependence, while the dependence on separation was modeled with powers of  $r^i$  with  $i = 0, 1, 2$ . The total number of broadband parameters being 48 (12 for each correlation).

Due to the large number of free parameters in this fit, we constrained the separation values to the range  $40 \text{ Mpc/h} < r < 180 \text{ Mpc/h}$  and applied extra priors to some of the nuisance parameters. The shift in the BAO parameters for this variation is  $\alpha_{\parallel} \sim 0.001$  and  $\alpha_{\perp} \sim -0.001$ , with a change in uncertainties that is negligible.

Similar to the variations in the correlation measurement, we did not detect any significant shift in the BAO parameters.

---

<sup>6</sup>C IV is the main contributor to the contamination produced by absorption from non-hydrogen lines. Its bias impacts the amplitude of the contamination.



## CONCLUSION

Modern cosmology is increasingly acquiring more and better observational data with the evolution of instrumental apparatus. In this context, the use of synthetic data or mock data has become a fundamental tool for any survey that wants to harness the full potential of its measurement devices.

In Chapter 5, we have presented CoLoRe, a parallelizable code that allows for the rapid generation of synthetic data with multiple tracers simultaneously in a coherent manner. Tools like this are crucial for testing pipelines, estimating covariances, and generating forecasts for future surveys. With the generation of more extensive and precise maps in a wide range of probes, the assistance of tools capable of generating multiple such probes simultaneously is highly valuable.

Nevertheless, it is vital to emphasize the limitations of CoLoRe. The execution speed is counterbalanced by the simplicity of the described fields, typically limited to Gaussian fields that rely on lognormal transformations to generate a physical non-linear density field.

One key advantage of CoLoRe is its modularity, offering a relatively straightforward way to improve its performance. We also improved quasar clustering for Lyman- $\alpha$  forest using Lagrangian Perturbation Theory (LPT). This progress is motivated by better clustering at small scales compared to the lognormal model, as well as properties such as the non-linear broadening of the Baryon Acoustic Oscillations (BAO) peak (Kirkby et al., 2013), which would help validate pipelines in a more realistic context. The working group is already working on generating Lyman- $\alpha$  forest mocks with LPT fields for the Dark Energy Spectroscopic Instrument (DESI) DR2 analysis, expecting to capitalize on these advantages.

Improvements in the synthetic data generated with CoLoRe would be greatly beneficial for future cosmological analyses. Specifically, for Lyman- $\alpha$  forest, having mocks with better clustering at small scales would enable analyses that use these smaller scales. This would allow to go beyond BAO and better constrain

the expansion of the universe (Cuceu et al., 2023b).

However, the core of my thesis work has been within the DESI Collaboration, focusing on Lyman- $\alpha$  science. My contribution to the collaboration has been comprehensive, involving both the generation of synthetic data with CoLoRe and the use of real data. The scientific outcome includes the publication of the Lyman- $\alpha$  forest fluctuation catalog for DESI Early Data Release (EDR), as well as the validation work for the Data Release 1 (DR1) BAO analysis.

The Lyman- $\alpha$  forest fluctuation catalog presented in Chapter 6, along with associated publications, has solidified Lyman- $\alpha$  as a stable measure of BAO at high redshift. The work on these publications helped adapt previously used methods, understand the new instrument and its peculiarities, untangle issues associated with synthetic data, and rectify minor errors in the data pipeline.

In this publication, we employed a weighting system to optimize correlation measurement, including an additional term in the weight expression (Eq. (6.10)). This improved the precision of auto-correlation measurements by 25%. However, there is still room for improvement, and the use of optimal weighting, where each pair of pixels is associated with a realistic covariance based on separations, is seen as the ideal scenario. This has already been correctly implemented for 1D (Karaçaylı et al., 2024), and could also be implemented in the future for 3D analyses.

The scale of modern surveys and the amount of data involved in their analyses require the development of complex pipelines that necessitate thorough validation to ensure the quality of the results. In Chapter 7, we described the validation with real data of the Lyman- $\alpha$  forest BAO analysis for the DESI DR1. The Lyman- $\alpha$  results from this analysis were an essential key in the BAO analysis for the survey DR1, significantly extending the redshift range that galaxy clustering can cover.

BAO has proven to be a powerful tool for constraining the clustering of galaxies. DESI, with just one year of observations, has been able to improve upon the measurements taken by the previous Baryon Oscillation Spectroscopic Survey (BOSS) and Extended Baryon Oscillation Spectroscopic Survey (eBOSS) surveys combined. The measurement of Lyman- $\alpha$  forest has extended this to a  $z_{\text{eff}} = 2.33$ , helping to obtain better cosmological constraints.

Although we have not yet been able to discover exactly what dark energy is, DESI has helped bring us closer to that moment by providing better constraints on cosmological parameters and indicating that its behavior might not exactly match that of a cosmological constant.

## BIBLIOGRAPHY

- Bessel, F. W. (1838). "On the parallax of 61 Cygni". In: *Mon. Not. Roy. Astron. Soc.* 4, pp. 152–161. DOI: [10.1093/mnras/4.17.152](https://doi.org/10.1093/mnras/4.17.152) (cit. on p. vii).
- Einstein, A. (1905). "Zur Elektrodynamik bewegter Körper". In: *Annalen der Physik* 322.10, pp. 891–921. DOI: [10.1002/andp.19053221004](https://doi.org/10.1002/andp.19053221004) (cit. on p. 2).
- Leavitt, H. S. (1908). "1777 variables in the Magellanic Clouds". In: *Annals of Harvard College Observatory* 60, pp. 87–108.3 (cit. on p. vii).
- Einstein, A. (1915). "Die Feldgleichungen der Gravitation". In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pp. 844–847 (cit. on pp. viii, 1).
- Hubble, E. P. (1925). "Cepheids in Spiral Nebulae". In: *Popular Astronomy* 33, pp. 252–255 (cit. on p. vii).
- Hubble, E. (1929). "A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae". In: *Proceedings of the National Academy of Science* 15.3, pp. 168–173. DOI: [10.1073/pnas.15.3.168](https://doi.org/10.1073/pnas.15.3.168) (cit. on pp. viii, 4, 6).
- Einstein, A. and W. de Sitter (1932). "On the Relation between the Expansion and the Mean Density of the Universe". In: *Proceedings of the National Academy of Science* 18.3, pp. 213–214. DOI: [10.1073/pnas.18.3.213](https://doi.org/10.1073/pnas.18.3.213) (cit. on p. viii).
- Bahcall, J. N. and E. E. Salpeter (1965). "On the Interaction of Radiation from Distant Sources with the Intervening Medium." In: *ApJ* 142, pp. 1677–1680. DOI: [10.1086/148460](https://doi.org/10.1086/148460) (cit. on p. 28).
- Dicke, R. H. et al. (1965). "Cosmic Black-Body Radiation." In: *ApJ* 142, pp. 414–419. DOI: [10.1086/148306](https://doi.org/10.1086/148306) (cit. on p. 24).
- Gunn, J. E. and B. A. Peterson (1965). "On the Density of Neutral Hydrogen in Intergalactic Space." In: *ApJ* 142, pp. 1633–1636. DOI: [10.1086/148444](https://doi.org/10.1086/148444) (cit. on pp. 28, 29).



- Penzias, A. A. and R. W. Wilson (1965). "A Measurement of Excess Antenna Temperature at 4080 Mc/s." In: *ApJ* 142, pp. 419–421. DOI: [10.1086/148307](https://doi.org/10.1086/148307) (cit. on p. 24).
- Scheuer, P. A. G. (1965). "A Sensitive Test for the Presence of Atomic Hydrogen in Intergalactic Space". In: *Nature* 207.5000, p. 963. DOI: [10.1038/207963a0](https://doi.org/10.1038/207963a0) (cit. on p. 28).
- Schmidt, M. (1965). "Large Redshifts of Five Quasi-Stellar Sources." In: *ApJ* 141, p. 1295. DOI: [10.1086/148217](https://doi.org/10.1086/148217) (cit. on p. 28).
- Sachs, R. K. and A. M. Wolfe (1967). "Perturbations of a Cosmological Model and Angular Variations of the Microwave Background". In: *ApJ* 147, p. 73. DOI: [10.1086/148982](https://doi.org/10.1086/148982) (cit. on pp. 24, 51).
- Sunyaev, R. A. and Y. B. Zeldovich (1970). "Small-Scale Fluctuations of Relic Radiation". In: *Annual Review of Astron and Astrophys* 7.1, pp. 3–19. DOI: [10.1007/BF00653471](https://doi.org/10.1007/BF00653471) (cit. on p. 24).
- Meszaros, P. (1974). "The behaviour of point masses in an expanding cosmological substratum." In: *Astron. Astrophys.* 37.2, pp. 225–228 (cit. on p. 17).
- Baldwin, J. A. (1977). "Luminosity Indicators in the Spectra of Quasi-Stellar Objects". In: *ApJ* 214, pp. 679–684. DOI: [10.1086/155294](https://doi.org/10.1086/155294) (cit. on p. 114).
- Peebles, P. J. E. (1980). *The large-scale structure of the universe* (cit. on p. 26).
- Sunyaev, R. A. and I. B. Zeldovich (1980). "Microwave background radiation as a probe of the contemporary structure and history of the universe". In: *Annual Review of Astron and Astrophys* 18, pp. 537–560. DOI: [10.1146/annurev.aa.18.090180.002541](https://doi.org/10.1146/annurev.aa.18.090180.002541) (cit. on p. 24).
- Hockney, R. W. and J. W. Eastwood (1981). *Computer Simulation Using Particles* (cit. on p. 48).
- Kaiser, N. (1984). "On the spatial correlations of Abell clusters." In: *Astrophys. J. Let.* 284, pp. L9–L12. DOI: [10.1086/184341](https://doi.org/10.1086/184341) (cit. on p. 25).
- Kaiser, N. (1987). "Clustering in real space and in redshift space". In: *Mon. Not. Roy. Astron. Soc.* 227, pp. 1–21. DOI: [10.1093/mnras/227.1.1](https://doi.org/10.1093/mnras/227.1.1) (cit. on p. 26).
- Kaiser, N. (1987). "Clustering in real space and in redshift space". In: *Monthly Notices of the Royal Astronomical Society* 227.1, pp. 1–21. ISSN: 0035-8711. DOI: [10.1093/mnras/227.1.1](https://doi.org/10.1093/mnras/227.1.1). eprint: <https://academic.oup.com/mnras/article-pdf/227/1/1/18522208/mnras227-0001.pdf>. URL: <https://doi.org/10.1093/mnras/227.1.1> (cit. on p. 65).
- Mather, J. C. et al. (1990). "A Preliminary Measurement of the Cosmic Microwave Background Spectrum by the Cosmic Background Explorer (COBE) Satellite". In: *Astrophys. J. Let.* 354, p. L37. DOI: [10.1086/185717](https://doi.org/10.1086/185717) (cit. on p. 24).

- Coles, P. and B. Jones (1991). “A lognormal model for the cosmological mass distribution.” In: *Mon. Not. Roy. Astron. Soc.* 248, pp. 1–13. DOI: [10.1093/mnras/248.1.1](#) (cit. on pp. 40, 45, 46).
- Smoot, G. F. et al. (1992). “Structure in the COBE Differential Microwave Radiometer First-Year Maps”. In: *Astrophys. J. Let.* 396, p. L1. DOI: [10.1086/186504](#) (cit. on p. 24).
- Zaroubi, S. and Y. Hoffman (1993). “Clustering in Redshift Space: Linear Theory”. In: *arXiv e-prints*, astro-ph/9311013, astro-ph/9311013. DOI: [10.48550/arXiv.astro-ph/9311013](#). arXiv: [astro-ph/9311013](#) [astro-ph] (cit. on p. 26).
- Hu, W. and M. White (1996). “Acoustic Signatures in the Cosmic Microwave Background”. In: *ApJ* 471, p. 30. DOI: [10.1086/177951](#). arXiv: [astro-ph/9602019](#) [astro-ph] (cit. on p. 18).
- Croft, R. A. C. et al. (1998). “Recovery of the Power Spectrum of Mass Fluctuations from Observations of the Ly $\alpha$  Forest”. In: *ApJ* 495.1, pp. 44–62. DOI: [10.1086/305289](#). arXiv: [astro-ph/9708018](#) [astro-ph] (cit. on p. 80).
- Riess, A. G. et al. (1998). “Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant”. In: *AJ* 116.3, pp. 1009–1038. DOI: [10.1086/300499](#). arXiv: [astro-ph/9805201](#) [astro-ph] (cit. on pp. viii, 4).
- Perlmutter, S. et al. (1999). “Measurements of  $\Omega$  and  $\Lambda$  from 42 High-Redshift Supernovae”. In: *ApJ* 517.2, pp. 565–586. DOI: [10.1086/307221](#). arXiv: [astro-ph/9812133](#) [astro-ph] (cit. on pp. viii, 4).
- McDonald, P. et al. (2000). “The Observed Probability Distribution Function, Power Spectrum, and Correlation Function of the Transmitted Flux in the Ly $\alpha$  Forest”. In: *ApJ* 543.1, pp. 1–23. DOI: [10.1086/317079](#). arXiv: [astro-ph/9911196](#) [astro-ph] (cit. on p. 80).
- Bartelmann, M. and P. Schneider (2001). “Weak gravitational lensing”. In: *Phys. Rept.* 340.4-5, pp. 291–472. DOI: [10.1016/S0370-1573\(00\)00082-X](#). arXiv: [astro-ph/9912508](#) [astro-ph] (cit. on p. 33).
- Carroll, S. M. (2001). “The Cosmological Constant”. In: *Living Reviews in Relativity* 4.1, 1, p. 1. DOI: [10.12942/lrr-2001-1](#). arXiv: [astro-ph/0004075](#) [astro-ph] (cit. on p. 2).
- Bernardeau, F. et al. (2002). “Large-scale structure of the Universe and cosmological perturbation theory”. In: *Phys. Rept.* 367.1-3, pp. 1–248. DOI: [10.1016/S0370-1573\(02\)00135-7](#). arXiv: [astro-ph/0112551](#) [astro-ph] (cit. on pp. 40, 47, 48).
- Schneider, P., L. van Waerbeke, and Y. Mellier (2002). “B-modes in cosmic shear from source redshift clustering”. In: *Astron. Astrophys.* 389, pp. 729–741.

- DOI: [10.1051/0004-6361:20020626](#). arXiv: [astro-ph/0112441](#) [[astro-ph](#)] (cit. on p. 64).
- Scoccimarro, R. and R. K. Sheth (2002). “PTHALOS: a fast method for generating mock galaxy distributions”. In: *Mon. Not. Roy. Astron. Soc.* 329.3, pp. 629–640. DOI: [10.1046/j.1365-8711.2002.04999.x](#). arXiv: [astro-ph/0106120](#) [[astro-ph](#)] (cit. on p. 40).
- Taffoni, G., P. Monaco, and T. Theuns (2002). “PINOCCHIO and the hierarchical build-up of dark matter haloes”. In: *Mon. Not. Roy. Astron. Soc.* 333.3, pp. 623–632. DOI: [10.1046/j.1365-8711.2002.05441.x](#). arXiv: [astro-ph/0109324](#) [[astro-ph](#)] (cit. on p. 40).
- Bennett, C. L. et al. (2003). “First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Foreground Emission”. In: *ApJS* 148.1, pp. 97–117. DOI: [10.1086/377252](#). arXiv: [astro-ph/0302208](#) [[astro-ph](#)] (cit. on p. 24).
- McDonald, P. (2003). “Toward a Measurement of the Cosmological Geometry at  $z \sim 2$ : Predicting  $\text{Ly}\alpha$  Forest Correlation in Three Dimensions and the Potential of Future Data Sets”. In: *ApJ* 585.1, pp. 34–51. DOI: [10.1086/345945](#). arXiv: [astro-ph/0108064](#) [[astro-ph](#)] (cit. on p. 30).
- Abdalla, F. B. and S. Rawlings (2005). “Probing dark energy with baryonic oscillations and future radio surveys of neutral hydrogen”. In: *Mon. Not. Roy. Astron. Soc.* 360.1, pp. 27–40. DOI: [10.1111/j.1365-2966.2005.08650.x](#). arXiv: [astro-ph/0411342](#) [[astro-ph](#)] (cit. on p. 56).
- Cole, S. et al. (2005). “The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications”. In: *Mon. Not. Roy. Astron. Soc.* 362.2, pp. 505–534. DOI: [10.1111/j.1365-2966.2005.09318.x](#). arXiv: [astro-ph/0501174](#) [[astro-ph](#)] (cit. on p. 46).
- Eisenstein, D. J. et al. (2005). “Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies”. In: *ApJ* 633.2, pp. 560–574. DOI: [10.1086/466512](#). arXiv: [astro-ph/0501171](#) [[astro-ph](#)] (cit. on p. 28).
- Górski, K. M. et al. (2005). “HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere”. In: *ApJ* 622.2, pp. 759–771. DOI: [10.1086/427976](#). arXiv: [astro-ph/0409513](#) [[astro-ph](#)] (cit. on pp. 36, 43).
- Gabasch, A. et al. (2006). “The evolution of the luminosity functions in the FORS deep field from low to high redshift. II. The red bands”. In: *Astron. Astrophys.* 448.1, pp. 101–121. DOI: [10.1051/0004-6361:20053986](#). arXiv: [astro-ph/0510339](#) [[astro-ph](#)] (cit. on p. 60).

- Lewis, A. and A. Challinor (2006). “Weak gravitational lensing of the CMB”. In: Phys. Rept. 429.1, pp. 1–65. DOI: [10.1016/j.physrep.2006.03.002](https://doi.org/10.1016/j.physrep.2006.03.002). arXiv: [astro-ph/0601594](https://arxiv.org/abs/astro-ph/0601594) [astro-ph] (cit. on p. 33).
- McDonald, P. et al. (2006). “The Ly $\alpha$  Forest Power Spectrum from the Sloan Digital Sky Survey”. In: *ApJS* 163.1, pp. 80–109. DOI: [10.1086/444361](https://doi.org/10.1086/444361). arXiv: [astro-ph/0405013](https://arxiv.org/abs/astro-ph/0405013) [astro-ph] (cit. on p. 103).
- Suzuki, N. (2006). “Quasar Spectrum Classification with Principal Component Analysis (PCA): Emission Lines in the Ly $\alpha$  Forest”. In: *ApJS* 163.1, pp. 110–121. DOI: [10.1086/499272](https://doi.org/10.1086/499272) (cit. on p. 107).
- Eisenstein, D. J., H.-J. Seo, and M. White (2007). “On the Robustness of the Acoustic Scale in the Low-Redshift Clustering of Matter”. In: *ApJ* 664.2, pp. 660–674. DOI: [10.1086/518755](https://doi.org/10.1086/518755). arXiv: [astro-ph/0604361](https://arxiv.org/abs/astro-ph/0604361) [astro-ph] (cit. on p. 17).
- Wilman, R. J. et al. (2008). “A semi-empirical simulation of the extragalactic radio continuum sky for next generation radio telescopes”. In: Mon. Not. Roy. Astron. Soc. 388.3, pp. 1335–1348. DOI: [10.1111/j.1365-2966.2008.13486.x](https://doi.org/10.1111/j.1365-2966.2008.13486.x). arXiv: [0805.3413](https://arxiv.org/abs/0805.3413) [astro-ph] (cit. on p. 67).
- Fixsen, D. J. (2009). “The Temperature of the Cosmic Microwave Background”. In: *ApJ* 707.2, pp. 916–920. DOI: [10.1088/0004-637X/707/2/916](https://doi.org/10.1088/0004-637X/707/2/916). arXiv: [0911.1955](https://arxiv.org/abs/0911.1955) [astro-ph.CO] (cit. on p. 23).
- LSST Science Collaboration et al. (2009). “LSST Science Book, Version 2.0”. In: *arXiv e-prints*, arXiv:0912.0201, arXiv:0912.0201. DOI: [10.48550/arXiv.0912.0201](https://doi.org/10.48550/arXiv.0912.0201). arXiv: [0912.0201](https://arxiv.org/abs/0912.0201) [astro-ph.IM] (cit. on pp. 25, 40).
- Bolton, A. S. and D. J. Schlegel (2010). “Spectro-Perfectionism: An Algorithmic Framework for Photon Noise-Limited Extraction of Optical Fiber Spectroscopy”. In: *PASP* 122.888, p. 248. DOI: [10.1086/651008](https://doi.org/10.1086/651008). arXiv: [0911.2689](https://arxiv.org/abs/0911.2689) [astro-ph.IM] (cit. on p. 36).
- Kitaura, F.-S., J. Jasche, and R. B. Metcalf (2010). “Recovering the non-linear density field from the galaxy distribution with a Poisson-lognormal filter”. In: Mon. Not. Roy. Astron. Soc. 403.2, pp. 589–604. DOI: [10.1111/j.1365-2966.2009.16163.x](https://doi.org/10.1111/j.1365-2966.2009.16163.x). arXiv: [0911.1407](https://arxiv.org/abs/0911.1407) [astro-ph.CO] (cit. on p. 46).
- Santos, M. G. et al. (2010). “Fast large volume simulations of the 21-cm signal from the reionization and pre-reionization epochs”. In: Mon. Not. Roy. Astron. Soc. 406.4, pp. 2421–2432. DOI: [10.1111/j.1365-2966.2010.16898.x](https://doi.org/10.1111/j.1365-2966.2010.16898.x). arXiv: [0911.2219](https://arxiv.org/abs/0911.2219) [astro-ph.CO] (cit. on p. 73).
- Beutler, F. et al. (2011). “The 6dF Galaxy Survey: baryon acoustic oscillations and the local Hubble constant”. In: Mon. Not. Roy. Astron. Soc. 416.4, pp. 3017–3032. DOI: [10.1111/j.1365-2966.2011.19250.x](https://doi.org/10.1111/j.1365-2966.2011.19250.x). arXiv: [1106.3366](https://arxiv.org/abs/1106.3366) [astro-ph.CO] (cit. on pp. 46, 47).

- Blake, C. et al. (2011). “The WiggleZ Dark Energy Survey: testing the cosmological model with baryon acoustic oscillations at  $z=0.6$ ”. In: *Mon. Not. Roy. Astron. Soc.* 415.3, pp. 2892–2909. DOI: [10.1111/j.1365-2966.2011.19077.x](https://doi.org/10.1111/j.1365-2966.2011.19077.x). arXiv: [1105.2862](https://arxiv.org/abs/1105.2862) [[astro-ph.CO](#)] (cit. on p. 46).
- Laureijs, R. et al. (2011). “Euclid Definition Study Report”. In: *arXiv e-prints*, arXiv:1110.3193, arXiv:1110.3193. DOI: [10.48550/arXiv.1110.3193](https://doi.org/10.48550/arXiv.1110.3193). arXiv: [1110.3193](https://arxiv.org/abs/1110.3193) [[astro-ph.CO](#)] (cit. on pp. 25, 40).
- Le Goff, J. M. et al. (2011). “Simulations of BAO reconstruction with a quasar Ly- $\alpha$  survey”. In: *Astron. Astrophys.* 534, A135, A135. DOI: [10.1051/0004-6361/201117736](https://doi.org/10.1051/0004-6361/201117736). arXiv: [1107.4233](https://arxiv.org/abs/1107.4233) [[astro-ph.CO](#)] (cit. on p. 46).
- Slosar, A. et al. (2011). “The Lyman- $\alpha$  forest in three dimensions: measurements of large scale flux correlations from BOSS 1st-year data”. In: *JCAP* 2011.9, 001, p. 001. DOI: [10.1088/1475-7516/2011/09/001](https://doi.org/10.1088/1475-7516/2011/09/001). arXiv: [1104.5244](https://arxiv.org/abs/1104.5244) [[astro-ph.CO](#)] (cit. on pp. 30, 80).
- Font-Ribera, A., P. McDonald, and J. Miralda-Escudé (2012a). “Generating mock data sets for large-scale Lyman- $\alpha$  forest correlation measurements”. In: *JCAP* 2012.1, 001, p. 001. DOI: [10.1088/1475-7516/2012/01/001](https://doi.org/10.1088/1475-7516/2012/01/001). arXiv: [1108.5606](https://arxiv.org/abs/1108.5606) [[astro-ph.CO](#)] (cit. on p. 46).
- Font-Ribera, A. et al. (2012b). “The large-scale cross-correlation of Damped Lyman alpha systems with the Lyman alpha forest: first measurements from BOSS”. In: *JCAP* 2012.11, 059, p. 059. DOI: [10.1088/1475-7516/2012/11/059](https://doi.org/10.1088/1475-7516/2012/11/059). arXiv: [1209.4596](https://arxiv.org/abs/1209.4596) [[astro-ph.CO](#)] (cit. on pp. 30, 109).
- Noterdaeme, P. et al. (2012). “Column density distribution and cosmological mass density of neutral gas: Sloan Digital Sky Survey-III Data Release 9”. In: *Astron. Astrophys.* 547, L1, p. L1. DOI: [10.1051/0004-6361/201220259](https://doi.org/10.1051/0004-6361/201220259). arXiv: [1210.1213](https://arxiv.org/abs/1210.1213) [[astro-ph.CO](#)] (cit. on p. 88).
- Seljak, U. (2012). “Bias, redshift space distortions and primordial nongaussianity of nonlinear transformations: application to Ly- $\alpha$  forest”. In: *JCAP* 2012.3, 004, p. 004. DOI: [10.1088/1475-7516/2012/03/004](https://doi.org/10.1088/1475-7516/2012/03/004). arXiv: [1201.0594](https://arxiv.org/abs/1201.0594) [[astro-ph.CO](#)] (cit. on p. 30).
- Addison, G. E., G. Hinshaw, and M. Halpern (2013). “Cosmological constraints from baryon acoustic oscillations and clustering of large-scale structure”. In: *Mon. Not. Roy. Astron. Soc.* 436.2, pp. 1674–1683. DOI: [10.1093/mnras/stt1687](https://doi.org/10.1093/mnras/stt1687). arXiv: [1304.6984](https://arxiv.org/abs/1304.6984) [[astro-ph.CO](#)] (cit. on p. 28).
- Busca, N. G. et al. (2013). “Baryon acoustic oscillations in the Ly $\alpha$  forest of BOSS quasars”. In: *Astron. Astrophys.* 552, A96, A96. DOI: [10.1051/0004-6361/201220724](https://doi.org/10.1051/0004-6361/201220724). arXiv: [1211.2616](https://arxiv.org/abs/1211.2616) [[astro-ph.CO](#)] (cit. on pp. 30, 80).



- Dawson, K. S. et al. (2013). “The Baryon Oscillation Spectroscopic Survey of SDSS-III”. In: *AJ* 145.1, 10, p. 10. DOI: [10.1088/0004-6256/145/1/10](https://doi.org/10.1088/0004-6256/145/1/10). arXiv: [1208.0022](https://arxiv.org/abs/1208.0022) [[astro-ph.CO](#)] (cit. on pp. viii, 30, 80).
- Kirkby, D. et al. (2013). “Fitting methods for baryon acoustic oscillations in the Lyman- $\alpha$  forest fluctuations in BOSS data release 9”. In: *JCAP* 2013.3, 024, p. 024. DOI: [10.1088/1475-7516/2013/03/024](https://doi.org/10.1088/1475-7516/2013/03/024). arXiv: [1301.3456](https://arxiv.org/abs/1301.3456) [[astro-ph.CO](#)] (cit. on pp. 30, 80, 123).
- Lee, K.-G. et al. (2013). “The BOSS Ly $\alpha$  Forest Sample from SDSS Data Release 9”. In: *AJ* 145.3, 69, p. 69. DOI: [10.1088/0004-6256/145/3/69](https://doi.org/10.1088/0004-6256/145/3/69). arXiv: [1211.5146](https://arxiv.org/abs/1211.5146) [[astro-ph.CO](#)] (cit. on pp. 30, 80).
- Levi, M. et al. (2013). “The DESI Experiment, a whitepaper for Snowmass 2013”. In: *arXiv e-prints*, arXiv:1308.0847, arXiv:1308.0847. arXiv: [1308.0847](https://arxiv.org/abs/1308.0847) [[astro-ph.CO](#)] (cit. on p. 80).
- Manera, M. et al. (2013). “The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: a large sample of mock galaxy catalogues”. In: *Mon. Not. Roy. Astron. Soc.* 428.2, pp. 1036–1054. DOI: [10.1093/mnras/sts084](https://doi.org/10.1093/mnras/sts084). arXiv: [1203.6609](https://arxiv.org/abs/1203.6609) [[astro-ph.CO](#)] (cit. on pp. 21, 40, 47, 73).
- Slosar, A. et al. (2013). “Measurement of baryon acoustic oscillations in the Lyman- $\alpha$  forest fluctuations in BOSS data release 9”. In: *JCAP* 2013.4, 026, p. 026. DOI: [10.1088/1475-7516/2013/04/026](https://doi.org/10.1088/1475-7516/2013/04/026). arXiv: [1301.3459](https://arxiv.org/abs/1301.3459) [[astro-ph.CO](#)] (cit. on pp. 30, 80, 103).
- Tassev, S., M. Zaldarriaga, and D. J. Eisenstein (2013). “Solving large scale structure in ten easy steps with COLA”. In: *JCAP* 2013.6, 036, p. 036. DOI: [10.1088/1475-7516/2013/06/036](https://doi.org/10.1088/1475-7516/2013/06/036). arXiv: [1301.0322](https://arxiv.org/abs/1301.0322) [[astro-ph.CO](#)] (cit. on pp. 40, 73).
- Alonso, D., P. G. Ferreira, and M. G. Santos (2014). “Fast simulations for intensity mapping experiments”. In: *Mon. Not. Roy. Astron. Soc.* 444.4, pp. 3183–3197. DOI: [10.1093/mnras/stu1666](https://doi.org/10.1093/mnras/stu1666). arXiv: [1405.1751](https://arxiv.org/abs/1405.1751) [[astro-ph.CO](#)] (cit. on pp. 39, 57–59).
- Font-Ribera, A. et al. (2014). “Quasar-Lyman  $\alpha$  forest cross-correlation from BOSS DR11: Baryon Acoustic Oscillations”. In: *JCAP* 2014.5, 027, p. 027. DOI: [10.1088/1475-7516/2014/05/027](https://doi.org/10.1088/1475-7516/2014/05/027). arXiv: [1311.1767](https://arxiv.org/abs/1311.1767) [[astro-ph.CO](#)] (cit. on pp. 30, 80).
- Kitaura, F. S., G. Yepes, and F. Prada (2014). “Modelling baryon acoustic oscillations with perturbation theory and stochastic halo biasing.” In: *Mon. Not. Roy. Astron. Soc.* 439, pp. L21–L25. DOI: [10.1093/mnrasl/slt172](https://doi.org/10.1093/mnrasl/slt172). arXiv: [1307.3285](https://arxiv.org/abs/1307.3285) [[astro-ph.CO](#)] (cit. on p. 40).
- White, M., J. L. Tinker, and C. K. McBride (2014). “Mock galaxy catalogues using the quick particle mesh method”. In: *Mon. Not. Roy. Astron. Soc.*

- 437.3, pp. 2594–2606. DOI: [10.1093/mnras/stt2071](https://doi.org/10.1093/mnras/stt2071). arXiv: [1309.5532](https://arxiv.org/abs/1309.5532) [[astro-ph.CO](#)] (cit. on p. 40).
- Arinyo-i-Prats, A. et al. (2015). “The non-linear power spectrum of the Lyman alpha forest”. In: *JCAP* 2015.12, pp. 017–017. DOI: [10.1088/1475-7516/2015/12/017](https://doi.org/10.1088/1475-7516/2015/12/017). arXiv: [1506.04519](https://arxiv.org/abs/1506.04519) [[astro-ph.CO](#)] (cit. on pp. 30, 121).
- Avila, S. et al. (2015). “HALOGEN: a tool for fast generation of mock halo catalogues”. In: *Monthly Notices of the Royal Astronomical Society* 450.2, 1856–1867. ISSN: 1365-2966. DOI: [10.1093/mnras/stv711](https://doi.org/10.1093/mnras/stv711). URL: <http://dx.doi.org/10.1093/mnras/stv711> (cit. on p. 40).
- Chuang, C.-H. et al. (2015). “EZmocks: extending the Zel’dovich approximation to generate mock galaxy catalogues with accurate clustering statistics”. In: *Mon. Not. Roy. Astron. Soc.* 446.3, pp. 2621–2628. DOI: [10.1093/mnras/stu2301](https://doi.org/10.1093/mnras/stu2301). arXiv: [1409.1124](https://arxiv.org/abs/1409.1124) [[astro-ph.CO](#)] (cit. on pp. 21, 40, 47).
- Delubac, T. et al. (2015). “Baryon acoustic oscillations in the Ly $\alpha$  forest of BOSS DR11 quasars”. In: *Astron. Astrophys.* 574, A59, A59. DOI: [10.1051/0004-6361/201423969](https://doi.org/10.1051/0004-6361/201423969). arXiv: [1404.1801](https://arxiv.org/abs/1404.1801) [[astro-ph.CO](#)] (cit. on pp. 30, 80, 103).
- Howlett, C., M. Manera, and W. J. Percival (2015). “L-PICOLA: A parallel code for fast dark matter simulation”. In: *Astron. Comput.* 12, pp. 109–126. DOI: [10.1016/j.ascom.2015.07.003](https://doi.org/10.1016/j.ascom.2015.07.003). arXiv: [1506.03737](https://arxiv.org/abs/1506.03737) [[astro-ph.CO](#)] (cit. on p. 40).
- Manera, M. et al. (2015). “The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: mock galaxy catalogues for the low-redshift sample”. In: *Mon. Not. Roy. Astron. Soc.* 447.1, pp. 437–445. DOI: [10.1093/mnras/stu2465](https://doi.org/10.1093/mnras/stu2465). arXiv: [1401.4171](https://arxiv.org/abs/1401.4171) [[astro-ph.CO](#)] (cit. on pp. 21, 40, 47).
- Pourtsidou, A. and R. B. Metcalf (2015). “Gravitational lensing of cosmological 21 cm emission”. In: *Mon. Not. Roy. Astron. Soc.* 448.3, pp. 2368–2383. DOI: [10.1093/mnras/stv102](https://doi.org/10.1093/mnras/stv102). arXiv: [1410.2533](https://arxiv.org/abs/1410.2533) [[astro-ph.CO](#)] (cit. on p. 57).
- Spergel, D. et al. (2015). “Wide-Field Infrared Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report”. In: *arXiv e-prints*, arXiv:1503.03757, arXiv:1503.03757. arXiv: [1503.03757](https://arxiv.org/abs/1503.03757) [[astro-ph.IM](#)] (cit. on p. 40).
- Abazajian, K. N. et al. (2016). “CMB-S4 Science Book, First Edition”. In: *arXiv e-prints*, arXiv:1610.02743, arXiv:1610.02743. arXiv: [1610.02743](https://arxiv.org/abs/1610.02743) [[astro-ph.CO](#)] (cit. on p. 40).
- Dawson, K. S. et al. (2016). “The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data”. In: *AJ* 151.2, 44, p. 44. DOI: [10.3847/0004-6256/151/2/44](https://doi.org/10.3847/0004-6256/151/2/44). arXiv: [1508.04473](https://arxiv.org/abs/1508.04473) [[astro-ph.CO](#)] (cit. on pp. viii, 30, 80).

- DESI Collaboration et al. (2016a). “The DESI Experiment Part I: Science, Targeting, and Survey Design”. In: *arXiv e-prints*, arXiv:1611.00036, arXiv:1611.00036. arXiv: [1611.00036 \[astro-ph.IM\]](#) (cit. on pp. ix, 40, 66, 80).
- (2016b). “The DESI Experiment Part II: Instrument Design”. In: *arXiv e-prints*, arXiv:1611.00037, arXiv:1611.00037. arXiv: [1611.00037 \[astro-ph.IM\]](#) (cit. on p. 35).
- Feng, Y. et al. (2016). “FASTPM: a new scheme for fast simulations of dark matter and haloes”. In: *Mon. Not. Roy. Astron. Soc.* 463.3, pp. 2273–2286. DOI: [10.1093/mnras/stw2123](#). arXiv: [1603.00476 \[astro-ph.CO\]](#) (cit. on p. 73).
- Izard, A., M. Crocce, and P. Fosalba (2016). “ICE-COLA: towards fast and accurate synthetic galaxy catalogues optimizing a quasi-N-body method”. In: *Monthly Notices of the Royal Astronomical Society* 459.3, 2327–2341. ISSN: 1365-2966. DOI: [10.1093/mnras/stw797](#). URL: <http://dx.doi.org/10.1093/mnras/stw797> (cit. on p. 40).
- McQuinn, M. (2016). “The Evolution of the Intergalactic Medium”. In: *Annual Review of Astron and Astrophys* 54, pp. 313–362. DOI: [10.1146/annurev-astro-082214-122355](#). arXiv: [1512.00086 \[astro-ph.CO\]](#) (cit. on p. 29).
- Newburgh, L. B. et al. (2016). “HIRAX: a probe of dark energy and radio transients”. In: *Ground-based and Airborne Telescopes VI*. Ed. by H. J. Hall, R. Gilmozzi, and H. K. Marshall. Vol. 9906. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 99065X, p. 99065X. DOI: [10.1117/12.2234286](#). arXiv: [1607.02059 \[astro-ph.IM\]](#) (cit. on p. 40).
- Silva, D. R. et al. (2016). “The Mayall z-band Legacy Survey”. In: *American Astronomical Society Meeting Abstracts #228*. Vol. 228. American Astronomical Society Meeting Abstracts, 317.02, p. 317.02 (cit. on p. 115).
- Xavier, H. S., F. B. Abdalla, and B. Joachimi (2016). “Improving lognormal models for cosmological fields”. In: *Mon. Not. Roy. Astron. Soc.* 459.4, pp. 3693–3710. DOI: [10.1093/mnras/stw874](#). arXiv: [1602.08503 \[astro-ph.CO\]](#) (cit. on p. 40).
- Agrawal, A. et al. (2017). “Generating log-normal mock catalog of galaxies in redshift space”. In: *JCAP* 2017.10, 003, p. 003. DOI: [10.1088/1475-7516/2017/10/003](#). arXiv: [1706.09195 \[astro-ph.CO\]](#) (cit. on p. 40).
- Bautista, J. E. et al. (2017). “Measurement of baryon acoustic oscillation correlations at  $z = 2.3$  with SDSS DR12  $\text{Ly}\alpha$ -Forests”. In: *Astron. Astrophys.* 603, A12, A12. DOI: [10.1051/0004-6361/201730533](#). arXiv: [1702.00176 \[astro-ph.CO\]](#) (cit. on pp. 30, 80, 103, 121).
- du Mas des Bourboux, H. et al. (2017). “Baryon acoustic oscillations from the complete SDSS-III  $\text{Ly}\alpha$ -quasar cross-correlation function at  $z = 2.4$ ”. In: *Astron.*



- Astrophys. 608, A130, A130. DOI: [10.1051/0004-6361/201731731](https://doi.org/10.1051/0004-6361/201731731). arXiv: [1708.02225](https://arxiv.org/abs/1708.02225) [astro-ph.CO] (cit. on pp. 30, 80).
- Izard, A., P. Fosalba, and M. Crocce (2017). “ICE-COLA: fast simulations for weak lensing observables”. In: *Monthly Notices of the Royal Astronomical Society* 473.3, 3051–3061. ISSN: 1365-2966. DOI: [10.1093/mnras/stx2544](https://doi.org/10.1093/mnras/stx2544). URL: <http://dx.doi.org/10.1093/mnras/stx2544> (cit. on p. 73).
- Villaescusa-Navarro, F., D. Alonso, and M. Viel (2017). “Baryonic acoustic oscillations from 21 cm intensity mapping: the Square Kilometre Array case”. In: *Mon. Not. Roy. Astron. Soc.* 466.3, pp. 2736–2751. DOI: [10.1093/mnras/stw3224](https://doi.org/10.1093/mnras/stw3224). arXiv: [1609.00019](https://arxiv.org/abs/1609.00019) [astro-ph.CO] (cit. on p. 59).
- Zou, H. et al. (2017). “Project Overview of the Beijing-Arizona Sky Survey”. In: *PASP* 129.976, p. 064101. DOI: [10.1088/1538-3873/aa65ba](https://doi.org/10.1088/1538-3873/aa65ba). arXiv: [1702.03653](https://arxiv.org/abs/1702.03653) [astro-ph.GA] (cit. on pp. 36, 115).
- Addison, G. E. et al. (2018). “Elucidating  $\Lambda$ CDM: Impact of Baryon Acoustic Oscillation Measurements on the Hubble Constant Discrepancy”. In: *ApJ* 853.2, 119, p. 119. DOI: [10.3847/1538-4357/aaa1ed](https://doi.org/10.3847/1538-4357/aaa1ed). arXiv: [1707.06547](https://arxiv.org/abs/1707.06547) [astro-ph.CO] (cit. on p. 28).
- Busca, N. and C. Balland (2018). “QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks”. In: *arXiv e-prints*, arXiv:1808.09955, arXiv:1808.09955. DOI: [10.48550/arXiv.1808.09955](https://doi.org/10.48550/arXiv.1808.09955). arXiv: [1808.09955](https://arxiv.org/abs/1808.09955) [astro-ph.IM] (cit. on p. 84).
- Desjacques, V., D. Jeong, and F. Schmidt (2018). “Large-scale galaxy bias”. In: *Phys. Rept.* 733, pp. 1–193. DOI: [10.1016/j.physrep.2017.12.002](https://doi.org/10.1016/j.physrep.2017.12.002). arXiv: [1611.09787](https://arxiv.org/abs/1611.09787) [astro-ph.CO] (cit. on p. 27).
- Font-Ribera, A., P. McDonald, and A. Slosar (2018). “How to estimate the 3D power spectrum of the Lyman- $\alpha$  forest”. In: *JCAP* 2018.1, 003, p. 003. DOI: [10.1088/1475-7516/2018/01/003](https://doi.org/10.1088/1475-7516/2018/01/003). arXiv: [1710.11036](https://arxiv.org/abs/1710.11036) [astro-ph.CO] (cit. on p. 103).
- Pérez-Ràfols, I. et al. (2018a). “The cosmological bias factor of damped Lyman alpha systems: dependence on metal line strength”. In: *Mon. Not. Roy. Astron. Soc.* 480.4, pp. 4702–4709. DOI: [10.1093/mnras/sty2158](https://doi.org/10.1093/mnras/sty2158). arXiv: [1805.00943](https://arxiv.org/abs/1805.00943) [astro-ph.GA] (cit. on p. 109).
- Pérez-Ràfols, I. et al. (2018b). “The SDSS-DR12 large-scale cross-correlation of damped Lyman alpha systems with the Lyman alpha forest”. In: *Mon. Not. Roy. Astron. Soc.* 473.3, pp. 3019–3038. DOI: [10.1093/mnras/stx2525](https://doi.org/10.1093/mnras/stx2525). arXiv: [1709.00889](https://arxiv.org/abs/1709.00889) [astro-ph.CO] (cit. on p. 109).
- Schaan, E., S. Ferraro, and D. N. Spergel (2018). “Weak lensing of intensity mapping: The cosmic infrared background”. In: *Phys. Rev. D* 97.12, 123539, p. 123539.

- DOI: [10.1103/PhysRevD.97.123539](https://doi.org/10.1103/PhysRevD.97.123539). arXiv: [1802.05706](https://arxiv.org/abs/1802.05706) [astro-ph.CO] (cit. on p. 57).
- The LSST Dark Energy Science Collaboration et al. (2018). “The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document”. In: *arXiv e-prints*, arXiv:1809.01669, arXiv:1809.01669. arXiv: [1809.01669](https://arxiv.org/abs/1809.01669) [astro-ph.CO] (cit. on p. 67).
- Ade, P. et al. (2019). “The Simons Observatory: science goals and forecasts”. In: *JCAP* 2019.2, 056, p. 056. DOI: [10.1088/1475-7516/2019/02/056](https://doi.org/10.1088/1475-7516/2019/02/056). arXiv: [1808.07445](https://arxiv.org/abs/1808.07445) [astro-ph.CO] (cit. on p. 40).
- Alonso, D. et al. (2019). “A unified pseudo- $C_\ell$  framework”. In: *Mon. Not. Roy. Astron. Soc.* 484.3, pp. 4127–4151. DOI: [10.1093/mnras/stz093](https://doi.org/10.1093/mnras/stz093). arXiv: [1809.09603](https://arxiv.org/abs/1809.09603) [astro-ph.CO] (cit. on p. 61).
- Blomqvist, M. et al. (2019). “Baryon acoustic oscillations from the cross-correlation of Ly $\alpha$  absorption and quasars in eBOSS DR14”. In: *Astron. Astrophys.* 629, A86, A86. DOI: [10.1051/0004-6361/201935641](https://doi.org/10.1051/0004-6361/201935641). arXiv: [1904.03430](https://arxiv.org/abs/1904.03430) [astro-ph.CO] (cit. on pp. 30, 80).
- de Sainte Agathe, V. et al. (2019). “Baryon acoustic oscillations at  $z = 2.34$  from the correlations of Ly $\alpha$  absorption in eBOSS DR14”. In: *Astron. Astrophys.* 629, A85, A85. DOI: [10.1051/0004-6361/201935638](https://doi.org/10.1051/0004-6361/201935638). arXiv: [1904.03400](https://arxiv.org/abs/1904.03400) [astro-ph.CO] (cit. on pp. 30, 80, 103).
- Dey, A. et al. (2019). “Overview of the DESI Legacy Imaging Surveys”. In: *AJ* 157.5, 168, p. 168. DOI: [10.3847/1538-3881/ab089d](https://doi.org/10.3847/1538-3881/ab089d). arXiv: [1804.08657](https://arxiv.org/abs/1804.08657) [astro-ph.IM] (cit. on p. 36).
- Guo, Z. and P. Martini (2019). “Classification of Broad Absorption Line Quasars with a Convolutional Neural Network”. In: *ApJ* 879.2, 72, p. 72. DOI: [10.3847/1538-4357/ab2590](https://doi.org/10.3847/1538-4357/ab2590). arXiv: [1901.04506](https://arxiv.org/abs/1901.04506) [astro-ph.GA] (cit. on p. 85).
- Stein, G., M. A. Alvarez, and J. R. Bond (2019). “The mass-Peak Patch algorithm for fast generation of deep all-sky dark matter halo catalogues and its N-body validation”. In: *Mon. Not. Roy. Astron. Soc.* 483.2, pp. 2236–2250. DOI: [10.1093/mnras/sty3226](https://doi.org/10.1093/mnras/sty3226). arXiv: [1810.07727](https://arxiv.org/abs/1810.07727) [astro-ph.CO] (cit. on p. 73).
- Witzemann, A. et al. (2019). “Simulated multitracer analyses with H I intensity mapping”. In: *Mon. Not. Roy. Astron. Soc.* 485.4, pp. 5519–5531. DOI: [10.1093/mnras/stz778](https://doi.org/10.1093/mnras/stz778). arXiv: [1808.03093](https://arxiv.org/abs/1808.03093) [astro-ph.CO] (cit. on p. 59).
- Allende Prieto, C. et al. (2020). “Preliminary Target Selection for the DESI Milky Way Survey (MWS)”. In: *Research Notes of the American Astronomical Society* 4.10, 188, p. 188. DOI: [10.3847/2515-5172/abc1dc](https://doi.org/10.3847/2515-5172/abc1dc). arXiv: [2010.11284](https://arxiv.org/abs/2010.11284) [astro-ph.GA] (cit. on p. 36).

- Cusin, G. et al. (2020). “Stochastic gravitational wave background anisotropies in the mHz band: astrophysical dependencies”. In: *Mon. Not. Roy. Astron. Soc.* 493.1, pp. L1–L5. DOI: [10.1093/mnrasl/slz182](https://doi.org/10.1093/mnrasl/slz182). arXiv: [1904.07757](https://arxiv.org/abs/1904.07757) [[astro-ph.CO](#)] (cit. on pp. xvi, 50, 51).
- Dodelson, S. and F. Schmidt (2020). *Modern Cosmology*. DOI: [10.1016/C2017-0-01943-2](https://doi.org/10.1016/C2017-0-01943-2) (cit. on p. 13).
- du Mas des Bourboux, H. et al. (2020). “The Completed SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations with Ly $\alpha$  Forests”. In: *ApJ* 901.2, 153, p. 153. DOI: [10.3847/1538-4357/abb085](https://doi.org/10.3847/1538-4357/abb085). arXiv: [2007.08995](https://arxiv.org/abs/2007.08995) [[astro-ph.CO](#)] (cit. on pp. xix, 30, 59, 80, 81, 85, 86, 88, 92, 96, 100, 102, 103, 105, 106, 109, 111, 120).
- Farr, J., A. Font-Ribera, and A. Pontzen (2020a). “Optimal strategies for identifying quasars in DESI”. In: *JCAP* 2020.11, 015, p. 015. DOI: [10.1088/1475-7516/2020/11/015](https://doi.org/10.1088/1475-7516/2020/11/015). arXiv: [2007.10348](https://arxiv.org/abs/2007.10348) [[astro-ph.CO](#)] (cit. on p. 84).
- Farr, J. et al. (2020b). “LyaCoLoRe: synthetic datasets for current and future Lyman- $\alpha$  forest BAO surveys”. In: *JCAP* 2020.3, 068, p. 068. DOI: [10.1088/1475-7516/2020/03/068](https://doi.org/10.1088/1475-7516/2020/03/068). arXiv: [1912.02763](https://arxiv.org/abs/1912.02763) [[astro-ph.CO](#)] (cit. on pp. 39, 55, 59).
- Karaçaylı, N. G., A. Font-Ribera, and N. Padmanabhan (2020). “Optimal 1D Ly  $\alpha$  forest power spectrum estimation - I. DESI-lite spectra”. In: *Mon. Not. Roy. Astron. Soc.* 497.4, pp. 4742–4752. DOI: [10.1093/mnras/staa2331](https://doi.org/10.1093/mnras/staa2331). arXiv: [2008.06421](https://arxiv.org/abs/2008.06421) [[astro-ph.CO](#)] (cit. on p. 103).
- Lyke, B. W. et al. (2020). “The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release”. In: *ApJS* 250.1, 8, p. 8. DOI: [10.3847/1538-4365/aba623](https://doi.org/10.3847/1538-4365/aba623). arXiv: [2007.09001](https://arxiv.org/abs/2007.09001) [[astro-ph.GA](#)] (cit. on p. 80).
- Nicola, A. et al. (2020). “Tomographic galaxy clustering with the Subaru Hyper Suprime-Cam first year public data release”. In: *JCAP* 2020.3, 044, p. 044. DOI: [10.1088/1475-7516/2020/03/044](https://doi.org/10.1088/1475-7516/2020/03/044). arXiv: [1912.08209](https://arxiv.org/abs/1912.08209) [[astro-ph.CO](#)] (cit. on p. 67).
- Pérez-Ràfols, I. et al. (2020). “Spectroscopic QUasar Extractor and redshift (z) Estimator SQUEZE - I. Methodology”. In: *Mon. Not. Roy. Astron. Soc.* 496.4, pp. 4931–4940. DOI: [10.1093/mnras/stz3467](https://doi.org/10.1093/mnras/stz3467). arXiv: [1903.00023](https://arxiv.org/abs/1903.00023) [[astro-ph.GA](#)] (cit. on p. 84).
- Planck Collaboration et al. (2020). “Planck 2018 results. VI. Cosmological parameters”. In: *Astron. Astrophys.* 641, A6, A6. DOI: [10.1051/0004-6361/201833910](https://doi.org/10.1051/0004-6361/201833910). arXiv: [1807.06209](https://arxiv.org/abs/1807.06209) [[astro-ph.CO](#)] (cit. on pp. 9, 18, 19, 24, 67).
- Raichoor, A. et al. (2020). “Preliminary Target Selection for the DESI Emission Line Galaxy (ELG) Sample”. In: *Research Notes of the American Astronomical*

- Society* 4.10, 180, p. 180. DOI: [10.3847/2515-5172/abc078](https://doi.org/10.3847/2515-5172/abc078). arXiv: [2010.11281](https://arxiv.org/abs/2010.11281) [[astro-ph.CO](#)] (cit. on p. 36).
- Ruiz-Macias, O. et al. (2020). “Preliminary Target Selection for the DESI Bright Galaxy Survey (BGS)”. In: *Research Notes of the American Astronomical Society* 4.10, 187, p. 187. DOI: [10.3847/2515-5172/abc25a](https://doi.org/10.3847/2515-5172/abc25a). arXiv: [2010.11283](https://arxiv.org/abs/2010.11283) [[astro-ph.GA](#)] (cit. on p. 36).
- Sinha, M. and L. H. Garrison (2020). “CORRFUNC - a suite of blazing fast correlation functions on the CPU”. In: *Mon. Not. Roy. Astron. Soc.* 491.2, pp. 3022–3041. DOI: [10.1093/mnras/stz3157](https://doi.org/10.1093/mnras/stz3157). arXiv: [1911.03545](https://arxiv.org/abs/1911.03545) [[astro-ph.CO](#)] (cit. on p. 63).
- Square Kilometre Array Cosmology Science Working Group et al. (2020). “Cosmology with Phase 1 of the Square Kilometre Array Red Book 2018: Technical specifications and performance forecasts”. In: *??jnlpASA* 37, e007, e007. DOI: [10.1017/pasa.2019.51](https://doi.org/10.1017/pasa.2019.51). arXiv: [1811.02743](https://arxiv.org/abs/1811.02743) [[astro-ph.CO](#)] (cit. on p. 40).
- Stein, G. et al. (2020). “The Websky extragalactic CMB simulations”. In: *JCAP* 2020.10, 012, p. 012. DOI: [10.1088/1475-7516/2020/10/012](https://doi.org/10.1088/1475-7516/2020/10/012). arXiv: [2001.08787](https://arxiv.org/abs/2001.08787) [[astro-ph.CO](#)] (cit. on p. 73).
- Yèche, C. et al. (2020). “Preliminary Target Selection for the DESI Quasar (QSO) Sample”. In: *Research Notes of the American Astronomical Society* 4.10, 179, p. 179. DOI: [10.3847/2515-5172/abc01a](https://doi.org/10.3847/2515-5172/abc01a). arXiv: [2010.11280](https://arxiv.org/abs/2010.11280) [[astro-ph.CO](#)] (cit. on p. 36).
- Zhou, R. et al. (2020). “Preliminary Target Selection for the DESI Luminous Red Galaxy (LRG) Sample”. In: *Research Notes of the American Astronomical Society* 4.10, 181, p. 181. DOI: [10.3847/2515-5172/abc0f4](https://doi.org/10.3847/2515-5172/abc0f4). arXiv: [2010.11282](https://arxiv.org/abs/2010.11282) [[astro-ph.CO](#)] (cit. on p. 36).
- Makiya, R., I. Kayo, and E. Komatsu (2021). “Ray-tracing log-normal simulation for weak gravitational lensing: application to the cross-correlation with galaxies”. In: *JCAP* 2021.3, 095, p. 095. DOI: [10.1088/1475-7516/2021/03/095](https://doi.org/10.1088/1475-7516/2021/03/095). arXiv: [2008.13195](https://arxiv.org/abs/2008.13195) [[astro-ph.CO](#)] (cit. on p. 40).
- DESI Collaboration et al. (2022). “Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument”. In: *AJ* 164.5, 207, p. 207. DOI: [10.3847/1538-3881/ac882b](https://doi.org/10.3847/1538-3881/ac882b). arXiv: [2205.10939](https://arxiv.org/abs/2205.10939) [[astro-ph.IM](#)] (cit. on p. 35).
- Ennesser, L. et al. (2022). “The impact and mitigation of broad-absorption-line quasars in Lyman  $\alpha$  forest correlations”. In: *Mon. Not. Roy. Astron. Soc.* 511.3, pp. 3514–3523. DOI: [10.1093/mnras/stac301](https://doi.org/10.1093/mnras/stac301). arXiv: [2111.09439](https://arxiv.org/abs/2111.09439) [[astro-ph.CO](#)] (cit. on pp. 81, 88, 118).
- Karaçaylı, N. G. et al. (2022). “Optimal 1D Ly  $\alpha$  forest power spectrum estimation - II. KODIAQ, SQUAD, and XQ-100”. In: *Mon. Not. Roy. Astron. Soc.*

- 509.2, pp. 2842–2855. DOI: [10.1093/mnras/stab3201](https://doi.org/10.1093/mnras/stab3201). arXiv: [2108.10870](https://arxiv.org/abs/2108.10870) [[astro-ph.CO](#)] (cit. on p. 103).
- Ramírez-Pérez, C. et al. (2022). “CoLoRe: fast cosmological realisations over large volumes with multiple tracers”. In: *JCAP* 2022.5, 002, p. 002. DOI: [10.1088/1475-7516/2022/05/002](https://doi.org/10.1088/1475-7516/2022/05/002). arXiv: [2111.05069](https://arxiv.org/abs/2111.05069) [[astro-ph.CO](#)] (cit. on pp. ix, 39).
- Alexander, D. M. et al. (2023). “The DESI Survey Validation: Results from Visual Inspection of the Quasar Survey Spectra”. In: *AJ* 165.3, 124, p. 124. DOI: [10.3847/1538-3881/acacfc](https://doi.org/10.3847/1538-3881/acacfc). arXiv: [2208.08517](https://arxiv.org/abs/2208.08517) [[astro-ph.GA](#)] (cit. on p. 84).
- Bailey et al. (2023). “Redrock: Spectroscopic Classification Pipeline for the Dark Energy Spectroscopic Instrument”. In: *in preparation* (cit. on p. 84).
- Bault, A. et al. (2023). “Impact of Redshift Errors on the 3D Cross-correlation of the Lyman- $\alpha$  Forest in DESI”. In: *in preparation* (cit. on p. 82).
- Brodzeller, A. et al. (2023). “Performance of the Quasar Spectral Templates for the Dark Energy Spectroscopic Instrument”. In: *AJ* 166.2, 66, p. 66. DOI: [10.3847/1538-3881/ace35d](https://doi.org/10.3847/1538-3881/ace35d). arXiv: [2305.10426](https://arxiv.org/abs/2305.10426) [[astro-ph.IM](#)] (cit. on pp. 84, 118).
- Chaussidon, E. et al. (2023). “Target Selection and Validation of DESI Quasars”. In: *ApJ* 944.1, 107, p. 107. DOI: [10.3847/1538-4357/acb3c2](https://doi.org/10.3847/1538-4357/acb3c2). arXiv: [2208.08511](https://arxiv.org/abs/2208.08511) [[astro-ph.CO](#)] (cit. on pp. 36, 80).
- Cooper, A. P. et al. (2023). “Overview of the DESI Milky Way Survey”. In: *ApJ* 947.1, 37, p. 37. DOI: [10.3847/1538-4357/acb3c0](https://doi.org/10.3847/1538-4357/acb3c0). arXiv: [2208.08514](https://arxiv.org/abs/2208.08514) [[astro-ph.GA](#)] (cit. on p. 36).
- Cuceu, A. et al. (2023a). “Constraints on the Cosmic Expansion Rate at Redshift 2.3 from the Lyman- $\alpha$  Forest”. In: *Phys. Rev. Lett.* 130.19, 191003, p. 191003. DOI: [10.1103/PhysRevLett.130.191003](https://doi.org/10.1103/PhysRevLett.130.191003). arXiv: [2209.13942](https://arxiv.org/abs/2209.13942) [[astro-ph.CO](#)] (cit. on p. 109).
- (2023b). “Constraints on the Cosmic Expansion Rate at Redshift 2.3 from the Lyman- $\alpha$  Forest”. In: *Phys. Rev. Lett.* 130.19, 191003, p. 191003. DOI: [10.1103/PhysRevLett.130.191003](https://doi.org/10.1103/PhysRevLett.130.191003). arXiv: [2209.13942](https://arxiv.org/abs/2209.13942) [[astro-ph.CO](#)] (cit. on p. 124).
- DESI Collaboration et al. (2023a). “The Early Data Release of the Dark Energy Spectroscopic Instrument”. In: *arXiv e-prints*, arXiv:2306.06308, arXiv:2306.06308. DOI: [10.48550/arXiv.2306.06308](https://doi.org/10.48550/arXiv.2306.06308). arXiv: [2306.06308](https://arxiv.org/abs/2306.06308) [[astro-ph.CO](#)] (cit. on pp. 37, 80, 81).
- DESI Collaboration et al. (2023b). “Validation of the Scientific Program for the Dark Energy Spectroscopic Instrument”. In: *arXiv e-prints*, arXiv:2306.06307,



- arXiv:2306.06307. DOI: [10.48550/arXiv.2306.06307](https://doi.org/10.48550/arXiv.2306.06307). arXiv: [2306.06307](https://arxiv.org/abs/2306.06307) [[astro-ph.CO](#)] (cit. on p. 80).
- Filbert, S. et al. (2023). “Broad Absorption Line Quasars in the Dark Energy Spectroscopic Instrument Early Data Release”. In: *arXiv e-prints*, arXiv:2309.03434, arXiv:2309.03434. DOI: [10.48550/arXiv.2309.03434](https://doi.org/10.48550/arXiv.2309.03434). arXiv: [2309.03434](https://arxiv.org/abs/2309.03434) [[astro-ph.CO](#)] (cit. on p. 85).
- Gontcho, S. et al. (2023). “Characterization of instrumental effects from DESI on Lyman- $\alpha$  3D correlations”. In: *in preparation* (cit. on p. 82).
- Gordon, C. et al. (2023). “3D correlations in the Lyman- $\alpha$  forest from early DESI data”. In: *JCAP* 2023.11, 045, p. 045. DOI: [10.1088/1475-7516/2023/11/045](https://doi.org/10.1088/1475-7516/2023/11/045). arXiv: [2308.10950](https://arxiv.org/abs/2308.10950) [[astro-ph.CO](#)] (cit. on pp. 32, 37, 80, 81, 102, 112, 119).
- Guy, J. et al. (2023). “The Spectroscopic Data Processing Pipeline for the Dark Energy Spectroscopic Instrument”. In: *AJ* 165.4, 144, p. 144. DOI: [10.3847/1538-3881/acb212](https://doi.org/10.3847/1538-3881/acb212). arXiv: [2209.14482](https://arxiv.org/abs/2209.14482) [[astro-ph.IM](#)] (cit. on pp. xx, 36, 84, 90, 93).
- Hahn, C. et al. (2023). “The DESI Bright Galaxy Survey: Final Target Selection, Design, and Validation”. In: *AJ* 165.6, 253, p. 253. DOI: [10.3847/1538-3881/accff8](https://doi.org/10.3847/1538-3881/accff8). arXiv: [2208.08512](https://arxiv.org/abs/2208.08512) [[astro-ph.CO](#)] (cit. on p. 36).
- Herrera, H. et al. (2023). “DESI Lyman- $\alpha$  synthetic spectra”. In: *in preparation* (cit. on pp. 37, 80, 82, 84).
- Karim, T. et al. (2023). “On the impact of the galaxy window function on cosmological parameter estimation”. In: *Mon. Not. Roy. Astron. Soc.* 525.1, pp. 311–324. DOI: [10.1093/mnras/stad2210](https://doi.org/10.1093/mnras/stad2210). arXiv: [2305.11956](https://arxiv.org/abs/2305.11956) [[astro-ph.CO](#)] (cit. on p. 20).
- Lan, T.-W. et al. (2023). “The DESI Survey Validation: Results from Visual Inspection of Bright Galaxies, Luminous Red Galaxies, and Emission-line Galaxies”. In: *ApJ* 943.1, 68, p. 68. DOI: [10.3847/1538-4357/aca5fa](https://doi.org/10.3847/1538-4357/aca5fa). arXiv: [2208.08516](https://arxiv.org/abs/2208.08516) [[astro-ph.CO](#)] (cit. on p. 84).
- Miller, T. N. et al. (2023). “The Optical Corrector for the Dark Energy Spectroscopic Instrument”. In: *arXiv e-prints*, arXiv:2306.06310, arXiv:2306.06310. DOI: [10.48550/arXiv.2306.06310](https://doi.org/10.48550/arXiv.2306.06310). arXiv: [2306.06310](https://arxiv.org/abs/2306.06310) [[astro-ph.IM](#)] (cit. on p. 35).
- Pérez-Ràfols, I. et al. (2023). “The cross-correlation of galaxies in absorption with the Lyman  $\alpha$  forest”. In: *Mon. Not. Roy. Astron. Soc.* 524.1, pp. 1464–1477. DOI: [10.1093/mnras/stad1994](https://doi.org/10.1093/mnras/stad1994). arXiv: [2210.02973](https://arxiv.org/abs/2210.02973) [[astro-ph.CO](#)] (cit. on p. 109).
- Raichoor, A. et al. (2023). “Target Selection and Validation of DESI Emission Line Galaxies”. In: *AJ* 165.3, 126, p. 126. DOI: [10.3847/1538-3881/acb213](https://doi.org/10.3847/1538-3881/acb213). arXiv: [2208.08513](https://arxiv.org/abs/2208.08513) [[astro-ph.CO](#)] (cit. on p. 36).

- Ravoux, C. et al. (2023). “The Dark Energy Spectroscopic Instrument: one-dimensional power spectrum from first Ly  $\alpha$  forest samples with Fast Fourier Transform”. In: *Monthly Notices of the Royal Astronomical Society* 526.4, pp. 5118–5140. ISSN: 0035-8711. DOI: [10.1093/mnras/stad3008](https://doi.org/10.1093/mnras/stad3008). eprint: <https://academic.oup.com/mnras/article-pdf/526/4/5118/52449716/stad3008.pdf>. URL: <https://doi.org/10.1093/mnras/stad3008> (cit. on pp. 32, 37, 81, 82, 97).
- Schlegel et al. (2023). In: *in preparation* (cit. on p. 36).
- Silber, J. H. et al. (2023). “The Robotic Multiobject Focal Plane System of the Dark Energy Spectroscopic Instrument (DESI)”. In: *AJ* 165.1, 9, p. 9. DOI: [10.3847/1538-3881/ac9ab1](https://doi.org/10.3847/1538-3881/ac9ab1). arXiv: [2205.09014](https://arxiv.org/abs/2205.09014) [astro-ph.IM] (cit. on p. 35).
- Zhou, R. et al. (2023). “Target Selection and Validation of DESI Luminous Red Galaxies”. In: *AJ* 165.2, 58, p. 58. DOI: [10.3847/1538-3881/aca5fb](https://doi.org/10.3847/1538-3881/aca5fb). arXiv: [2208.08515](https://arxiv.org/abs/2208.08515) [astro-ph.CO] (cit. on p. 36).
- Zou, J. et al. (2023). “The DESI Damped Ly $\alpha$  System Survey: Data Release”. In: *in preparation* (cit. on p. 85).
- Bovy, J. (2024). *Dynamics and Astrophysics of Galaxies*. Online graduate textbook. URL: <https://galaxiesbook.org/> (cit. on pp. xv, 33).
- Cuceu, A. et al. (2024). “Validation of the DESI 2024 Ly $\alpha$  forest BAO analysis using synthetic datasets”. In: *arXiv e-prints*, arXiv:2404.03004, arXiv:2404.03004. DOI: [10.48550/arXiv.2404.03004](https://doi.org/10.48550/arXiv.2404.03004). arXiv: [2404.03004](https://arxiv.org/abs/2404.03004) [astro-ph.CO] (cit. on p. 119).
- DES Collaboration et al. (2024). “Dark Energy Survey: A 2.1% measurement of the angular Baryonic Acoustic Oscillation scale at redshift  $z_{\text{eff}}=0.85$  from the final dataset”. In: *arXiv e-prints*, arXiv:2402.10696, arXiv:2402.10696. DOI: [10.48550/arXiv.2402.10696](https://doi.org/10.48550/arXiv.2402.10696). arXiv: [2402.10696](https://arxiv.org/abs/2402.10696) [astro-ph.CO] (cit. on p. 28).
- DESI Collaboration et al. (2024a). “DESI 2024 III: Baryon Acoustic Oscillations from Galaxies and Quasars”. In: *arXiv e-prints*, arXiv:2404.03000, arXiv:2404.03000. DOI: [10.48550/arXiv.2404.03000](https://doi.org/10.48550/arXiv.2404.03000). arXiv: [2404.03000](https://arxiv.org/abs/2404.03000) [astro-ph.CO] (cit. on pp. 28, 37, 111).
- DESI Collaboration et al. (2024b). “DESI 2024 IV: Baryon Acoustic Oscillations from the Lyman Alpha Forest”. In: *arXiv e-prints*, arXiv:2404.03001, arXiv:2404.03001. DOI: [10.48550/arXiv.2404.03001](https://doi.org/10.48550/arXiv.2404.03001). arXiv: [2404.03001](https://arxiv.org/abs/2404.03001) [astro-ph.CO] (cit. on pp. ix, xv, xxiii, 30, 31, 37, 111–113, 119–121).
- DESI Collaboration et al. (2024c). “DESI 2024 VI: Cosmological Constraints from the Measurements of Baryon Acoustic Oscillations”. In: *arXiv e-prints*,

- arXiv:2404.03002, arXiv:2404.03002. DOI: [10.48550/arXiv.2404.03002](https://doi.org/10.48550/arXiv.2404.03002). arXiv: [2404.03002](https://arxiv.org/abs/2404.03002) [[astro-ph.CO](#)] (cit. on p. 111).
- Etourneau, T. et al. (2024). “Mock data sets for the Eboss and DESI Lyman- $\alpha$  forest surveys”. In: *JCAP* 2024.5, 077, p. 077. DOI: [10.1088/1475-7516/2024/05/077](https://doi.org/10.1088/1475-7516/2024/05/077). arXiv: [2310.18996](https://arxiv.org/abs/2310.18996) [[astro-ph.CO](#)] (cit. on p. 112).
- Karaçaylı, N. G. et al. (2024). “Optimal 1D Ly  $\alpha$  forest power spectrum estimation - III. DESI early data”. In: *Mon. Not. Roy. Astron. Soc.* 528.3, pp. 3941–3963. DOI: [10.1093/mnras/stae171](https://doi.org/10.1093/mnras/stae171). arXiv: [2306.06316](https://arxiv.org/abs/2306.06316) [[astro-ph.CO](#)] (cit. on pp. 32, 37, 81, 82, 103, 124).
- Maus, M. et al. (2024). “A comparison of effective field theory models of redshift space galaxy power spectra for DESI 2024 and future surveys”. In: *arXiv e-prints*, arXiv:2404.07272, arXiv:2404.07272. DOI: [10.48550/arXiv.2404.07272](https://doi.org/10.48550/arXiv.2404.07272). arXiv: [2404.07272](https://arxiv.org/abs/2404.07272) [[astro-ph.CO](#)] (cit. on p. 27).
- Ramírez-Pérez, C. et al. (2024). “The Lyman- $\alpha$  forest catalogue from the Dark Energy Spectroscopic Instrument Early Data Release”. In: *Mon. Not. Roy. Astron. Soc.* 528.4, pp. 6666–6679. DOI: [10.1093/mnras/stad3781](https://doi.org/10.1093/mnras/stad3781). arXiv: [2306.06312](https://arxiv.org/abs/2306.06312) [[astro-ph.CO](#)] (cit. on pp. ix, 32, 37, 79).









2024

Analysis validation of the Lyman- $\alpha$  forest measurements from the Dark Energy Spectroscopic Instrument

César Ramírez Pérez

