Improving the Cosmic Statistics of Neutral Hydrogen

Naim Goksel Karacayli
Yale University Graduate School of Arts and Sciences, ngokselk@gmail.com

Follow this and additional works at: https://elischolar.library.yale.edu/gsas_dissertations

Recommended Citation
Karacayli, Naim Goksel, "Improving the Cosmic Statistics of Neutral Hydrogen" (2021). Yale Graduate School of Arts and Sciences Dissertations. 356. https://elischolar.library.yale.edu/gsas_dissertations/356
Abstract

Improving the Cosmic Statistics of Neutral Hydrogen

Naim Göksel Karaçaylı

2021

The study of the cosmological large-scale structure seeks to understand the makings and the evolution of the universe. In this subject, I worked on improving current techniques and their application to the existing large, high-precision cosmological data sets. Specifically, my dissertation explores boosting power spectrum measurements at large scales for 21-cm intensity maps through reconstruction, and at small scales for Ly$\alpha$ forest by developing and applying the optimal estimator to hundreds of high-resolution spectra.

The cosmic tidal reconstruction is a novel technique for low redshift ($z < 2$) 21-cm intensity mapping surveys (e.g. CHIME and HIRAX) that recovers the lost large-scale line-of-sight signal from local small-scale anisotropies formed by tidal interactions. My thesis shows this algorithm is robust against redshift space distortions and can recover the signal with approximately 70% efficiency for $k < 0.1 \, h \, \text{Mpc}^{-1}$ using N-body simulations. If angular modes are also lost (known as the foreground wedge), the efficiency drops down to 30–50% range. I further introduce an analytical framework based on perturbation theory, which correctly predicts the shape of the reconstructed field’s 2D power spectrum and reveals that the reconstruction mostly utilizes angular modes with $k > 0.3 \, h \, \text{Mpc}^{-1}$.

Through absorption lines in quasar spectra, the Ly$\alpha$ forest technique can probe matter in vast volumes far into the past ($2 \lesssim z \lesssim 5$) and at smaller scales than galaxy surveys ($r \lesssim 1 \, \text{Mpc}$). The power spectrum at these scales is shaped by the thermal state of the intergalactic medium (IGM), reionization history of the universe, neutrino masses and the nature of the dark matter. 1D power spectrum $P_{1D}$ has emerged as a competitive framework to study new physics, but has come with various challenges and systematic errors in analysis.
I implement the optimal quadratic estimator for $P_{1\text{D}}$ and generate synthetic spectra based on The Dark Energy Spectroscopic Instrument (DESI) specifications. Using these mock spectra, I prove robustness against relevant problems including quasar continuum errors and gaps in spectra due to bad pixels or masked high column density absorbers, show that an input fiducial power spectrum improves the accuracy, and provide simple 5-yr forecasts for DESI $P_{1\text{D}}$ measurements. I also apply the optimal estimator to the largest number of high-resolution, high-S/N spectra, obtained by combining Keck Observatory Database of Ionized Absorption toward Quasars (KODIAQ), The Spectral Quasar Absorption Database (SQUAD) and XQ-100 data sets. This project yields the most precise $P_{1\text{D}}$ measurement at small scales and should improve the mass of warm dark matter constraints by more than a factor of 2.
Improving the Cosmic Statistics of Neutral Hydrogen

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Naim Göksel Karaçaylı

Dissertation Director: Nikhil Padmanabhan

December 2021
Copyright © 2021 by Naim Göksel Karaçaylı
All rights reserved.
Acknowledgments

This thesis would not be possible without the enormous moral support I was fortunate to have over the six years from family and friends. I am thankful to have continuous support from my mother Berna Karaçaylı, my father Cemil Karaçaylı and my sister Aslı Karaçaylı.

I was yet lucky to have another family here in Connecticut. From the moment I set foot in the US, Süheyl Keskin helped me in many ways. I am grateful for all the time I spent with him, his family and friends. I thank Sadık Aydınlı, Cafer Keskin and his wife Ramona, Oktay Tartıcı and his wife Aylin, and everyone I met in Sunny Daes ice cream.

The beautiful friendships I developed during the graduate school made me the person I wanted to be. I would like to acknowledge my dear friends Mehmet and Ceyda Doğan; and especially thank Mehmet for introducing me to the labor movement. We spent meaningful and fun times with our small, like-minded, but fluctuating Turkish group where Ferzan Tapramaz, Hasan Yılmaz, Soner Albayrak and Kerim Gökarslan were members at different, sometimes overlapping, periods in time. Another special thanks to my dear friend Josh Stanley who pushed me when I needed and helped me overcome many personal problems. Without his consistent friendship I would not be the person I am today. I will remember our time during and after Local 33 elections as one of the most meaningful periods of my life. Finally, I would like to acknowledge some of my fellow organizers in this journey. I thank Judith Hoeller, Qing Xia, Lena Eckert-Erdheim, Charles Decker and Lukas Moe.
Contents

1 Introduction 1

1.1 Overview of Thesis 2
1.2 Theoretical Preliminaries 3
  1.2.1 The optimal estimator 5
1.3 The Intergalactic Medium 6
1.4 Observations & Key Projects 9
  1.4.1 21-cm intensity mapping surveys 10
  1.4.2 Lyman-alpha Forest 11

2 Tidal Reconstruction & 21-cm Intensity Maps 14

2.1 Introduction 14
2.2 Theory 16
  2.2.1 Tidal Interactions & Local Power Spectrum 17
  2.2.2 Estimators 21
2.3 Implementation 23
  2.3.1 Simulations 23
  2.3.2 Method 24
2.4 Results 25
  2.4.1 Tests on the Reconstruction 25
  2.4.2 Foregrounds 30
2.5 Theoretical Expectations ........................................ 32
  2.5.1 Results ......................................................... 36
2.6 Summary ............................................................. 39

3 Optimal 1D Lyman-alpha Forest Power Spectrum Estimator 43
  3.1 Introduction ....................................................... 43
  3.2 Method ............................................................. 46
    3.2.1 Ly \( \alpha \) Forest Specifics ................................. 47
    3.2.2 Continuum Limit .............................................. 49
  3.3 Implementation .................................................. 50
    3.3.1 Algorithm .................................................... 50
    3.3.2 Fiducial Power Spectrum ................................. 52
    3.3.3 Convergence .................................................. 53
    3.3.4 Validation .................................................... 53
  3.4 DESI-Lite Spectra ................................................. 60
    3.4.1 Gaps in Spectra .............................................. 64
    3.4.2 Continuum Marginalization ............................... 64
    3.4.3 Choice of Fiducial .......................................... 66
    3.4.4 Fisher Matrix Approximation ............................. 68
    3.4.5 5-Year Forecasts ............................................ 69
  3.5 Discussion ....................................................... 70
  3.6 Summary ............................................................. 72

4 Applying the Optimal Estimator to High-Resolution Quasar Spectra 74
  4.1 Introduction ....................................................... 74
  4.2 Data ............................................................... 76
    4.2.1 KODIAQ ....................................................... 76
    4.2.2 SQUAD .......................................................... 77
4.2.3 XQ-100 ............................................. 77
4.3 Method ............................................. 78
  4.3.1 Summary of the pipeline ....................... 78
  4.3.2 Mean flux ..................................... 80
  4.3.3 Quadratic estimator .......................... 82
  4.3.4 Validation ..................................... 86
4.4 Results ............................................ 91
  4.4.1 Consistency between data sets ................ 94
  4.4.2 Final combined results ....................... 95
  4.4.3 Damped Lyman-alpha absorbers ............... 96
  4.4.4 Continuum ................................... 98
  4.4.5 Side band power ............................. 100
  4.4.6 Systematic error budget ..................... 106
4.5 Discussion ....................................... 110
  4.5.1 Statistical power of our results ............. 110
  4.5.2 Remaining challenges ....................... 113
4.6 Summary ......................................... 114

5 Future Applications ............................ 116
  5.1 Applying the optimal estimator to DESI ........ 116
    5.1.1 Implementing the resolution matrix .......... 117
    5.1.2 Test and improve continuum marginalization 117
    5.1.3 Preliminary results from Survey Validation 117
    5.1.4 Realistic DESI mocks ....................... 119
  5.2 Tidal reconstruction for biased tracers .......... 121
    5.2.1 Photometric surveys ....................... 121
    5.2.2 Using spherical coordinates ............... 122
A  Tidal Reconstruction  123
   A.1  Time Dependent Functions  . . . . . . . . . . . . . . . . . . . . . 123
   A.2  Details on Reconstruction Theory  . . . . . . . . . . . . . . . . . . . 124
   A.3  Smoothing  . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125

B  Lyman-alpha Forest  127
   B.1  Saddle Point Approximation for the Mean Flux of Log-normal Mocks  . . 127

Bibliography  128
List of Figures

1.1 Simulations and observations of large-scale structure. 2
1.2 Diagram for reionization 7
1.3 Spectrum of SDSS J114308.87+345222.2 11

2.1 The reconstructed cross-correlation coefficients $r_{\kappa\delta}(k)$ in real and redshift spaces, and for multiple $k_\perp$ cutoffs. 26
2.2 The reconstructed cross-correlation coefficients $r_{\kappa\delta}(k)$ at $z = 1$ in redshift space after foreground removal with wedge. 27
2.3 The reconstructed cross-correlation coefficients $r(k_\perp, k_\parallel)$ in 2D in redshift space. 28
2.4 Figure 2.2, but $r_{\kappa\delta}(k_\perp, k_\parallel = 0)$. 29
2.5 Reconstructed power spectrum $P_\kappa$ of one simulation run compared to the theoretical estimate. 37
2.6 The fraction of cross correlation signal $\alpha(q, \theta)$ that can be achieved using modes up to $q$ and angle $\theta$. 39

3.1 The relative error in our measurements in the absence of any systematic. 54
3.2 Power spectrum estimates from 800,000 DESI-lite spectra. 55
3.2 Power spectrum estimates from 800,000 DESI-lite spectra continued. 56
3.3 Percentage of time spent in different steps for 800,000 DESI-lite spectra. 58
3.4 The normalized Fisher matrix between redshift bins. ..................... 61
3.5 The effect of gaps in \( P_{1D} \). ............................................. 63
3.6 The effect of continuum errors in \( P_{1D} \). ................................. 65
3.7 Comparing \( \chi^2 \) for three different fiducials. ............................ 67

4.1 Mean flux measurement of each data set when DLAs are masked. ...... 81
4.2 Signal-to-noise of each data set. ............................................. 82
4.3 Power spectrum estimates from one mock set. Error bars are replaced with bootstrap estimates. .................................................. 85
4.4 Proof of unbiased results from the average of 100 mocks. ................. 87
4.5 Noise power in one mock data. ................................................. 88
4.6 \( \chi^2 \) validation on 100 mock sets using the Fisher matrices from QMLE and 25,000 bootstrap realizations. ....................................... 88
4.7 The normalized bootstrap covariance matrix of data for \( k = 0.013 \text{ s} \text{ km}^{-1} \) bin. ................................................................. 90
4.8 The raw power spectrum measurements of KODIAQ, SQUAD and XQ-100 when DLAs are masked. ............................................. 92
4.9 Our raw and side band subtracted combined power spectrum measurements when DLAs are masked. ......................................... 93
4.10 The ratio of the measured power spectrum from the combined data set between DLAs kept and DLAs masked. ................................. 97
4.11 The ratio of the measured power spectrum without continuum marginalization to with marginalization. .................................... 98
4.12 Power spectrum at side band 1 with FFT and QMLE. ...................... 101
4.13 Power spectrum at side band 2 with FFT and QMLE. ...................... 102
4.14 Comparison of both side band power spectra. ................................ 103
4.15 The power spectra measured from the two side bands at \( z = 3.0 \). ...... 104
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.16</td>
<td>The residual correlation functions for both side bands.</td>
<td>105</td>
</tr>
<tr>
<td>4.17</td>
<td>Systematic error estimates with respect to statistical ones.</td>
<td>107</td>
</tr>
<tr>
<td>4.18</td>
<td>The suppression in $P_{1D}$ due to warm dark matter.</td>
<td>112</td>
</tr>
<tr>
<td>5.1</td>
<td>Preliminary DESI SV results</td>
<td>118</td>
</tr>
<tr>
<td>5.2</td>
<td>A QUICKQUASARS simulated spectrum for a quasar at $z = 2.4$.</td>
<td>120</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Fitting results for $P_{1D}$. .................................................. 52
3.2 Our model power spectrum and precision forecasts for 5-year DESI survey. 69
3.3 Fisher matrix forecasts for 5-year DESI survey for different spectral qualities. 70
4.1 Best-fitted parameters to our mean flux measurements of each data set when DLAs are masked. ........................................... 81
4.2 Doublets used in SB correlation function analysis. ...................... 106
4.3 Comparison of different data sets’ statistical warm dark matter mass con-
straining power. ................................................................. 113
Chapter 1

Introduction

Discovering the accelerated expansion of our homogeneous universe presented one major mystery to modern science—dark energy. Even though the cosmological constant is consistent with current observations, a diverse set of ideas such as modified gravity theories, new particles and fields has been relentlessly investigated to better understand its nature. To this purpose, the physics beyond homogeneous universe, that is, the physics of structure formation in the universe is a prominent field.

Inhomogeneity in the early universe eventually forms stars and planets and notably our planet—Earth. Yet, we do not need to look at small objects such as Earth, our Solar system or our own Milky Way to see this feature. Manifested as the web-like pattern in observations\(^1\) and reproduced by numerical simulations\(^2\) as shown on figure 1.1, inhomogeneity permeates the universe even on large-scales. By quantifying this pattern as deviations from the mean, we can study these structures and differentiate between cosmological models.

The standard model of cosmology (ΛCDM) is a homogeneous and isotropic universe where structures spring from density fluctuations. Uniform, randomly distributed particles, a simple first guess, do not describe our universe. Even though they form an almost homogeneous and isotropic system, these particles are uncorrelated by construction; it is

\(^{1}\)http://www.sdss.org/science/orangepie/
\(^{2}\)https://www.cosmosim.org/cms/images-and-movies/
neither more nor less likely to find particles near other particles. However, gravity causes matter to collapse, cluster and correlate in denser regions. The properties of this clustering give us the most-needed clues to the nature of the dark matter and dark energy.

1.1 Overview of Thesis

My thesis is part of advancing the precision cosmology as large and precise data sets become available. In it, I studied and applied certain theoretical techniques to improve the clustering measurements of the neutral hydrogen. This thesis can be categorized under two H I transition lines. Chapter 2 is a published work that reviews and advances a proposed theoretical reconstruction algorithm to improve low-redshift, 21-cm emission intensity mapping data (Pen et al., 2012; Zhu et al., 2016; Zhu et al., 2018a). Specifically, I performed further tests with more realistic data loss and in redshift space; and I introduced an analytic framework to understand and forecast the performance of the reconstruction. The remaining two chapters study the Lyman-alpha absorption features in quasar spectra. Chapter 3 is another published work in which I implemented the theoretically optimal
method to investigate these features. The optimal quadratic estimator of the 1D power spectrum (QMLE) has been sparsely applied to real data—the only two publications are McDonald et al. (2006) and Palanque-Delabrouille et al. (2013). I developed an improved version of QMLE and performed extensive tests on synthetic spectra that resembles the future Dark Energy Spectroscopic Instrument data. In Chapter 4, I applied this estimator to the largest high-resolution, high-S/N data. This work obtained the most precise power spectrum measurements at small scales, and I forecast it to improve warm dark matter mass constraints by more than a factor of 2. Chapter 5 discusses some future applications of this thesis.

1.2 Theoretical Preliminaries

The structures formed by gravity is encoded in the density field $\rho(x)$. The scales of interest are large enough to assume an almost homogeneous universe, so it is best to study the deviations from the average density $\bar{\rho}$. We define the over-density field $\delta(x) = \rho(x)/\bar{\rho} - 1$. The correlation function captures the statistical feature of structures in the density field.

$$\xi(r) = \langle \delta(x)\delta(x + r) \rangle$$ (1.1)

The isotropy and homogeneity reduces its dependence to the distance between two points, independent of location and direction. Moreover, the early universe fluctuations are assumed to be a Gaussian random field; and the correlation function fully describes such a field. In general, we can define higher-order statistics such as three- and four-point functions.

We can also examine the field in Fourier space. I assume the following Fourier conven-
These Fourier modes are uncorrelated in the linear regime. The two-point function in Fourier space is the power spectrum. Isotropy constrains its dependence to $k = |k|$. The three-point function is called the bispectrum.

$$\langle \delta(k)\delta(k') \rangle = (2\pi)^3 \delta_D(k + k') P(k) \quad (1.4)$$
$$\langle \delta(k_1)\delta(k_2)\delta(k_3) \rangle = (2\pi)^3 \delta_D(k_1 + k_2 + k_3) B(k_1, k_2) \quad (1.5)$$

These ideal conditions (Gaussianity and linearity) will be violated in reality as expected. Because gravity is non-linear, different modes are coupled to each other. There is also an observational effect: We relate a measured redshift to the line-of-sight distance by a given Hubble expansion. Yet, galaxies have peculiar velocities in addition to Hubble expansion since they fall towards over-densities and form structures. The line-of-sight positions are hence mismeasured because of these peculiar velocities. Specifically, the new distorted positions will be $s = x + v_{1\parallel} \hat{x}_{\parallel} / H(z)$, where $H(z)$ is the Hubble parameter at redshift $z$. The structure growth rate is imprinted in the resulting redshift-space anisotropic clustering, which can then probe deviations from general relativity. Having theoretical tools to understand these effects is important to test different models and to have enough sensitivity in any analysis. Combining linear regime and plane-parallel approximation yields a simple relation between distorted power spectrum $P^{(s)}(k)$ and undistorted power spectrum $P(k)$ (Kaiser, 1987).

$$P^{(s)}(k) = \left(1 + f_{\mu}^2\right)^2 P(k), \quad (1.6)$$
where $\mu = k_z/k$ and $f \approx \Omega_m^{0.55}$ is the linear growth factor\(^3\). However, more advanced models exist (Scoccimarro, 2004; Matsubara, 2008a,b).

### 1.2.1 The optimal estimator

An important recurrent theme of this thesis is the optimal estimator. This estimator is also known the optimal quadratic estimator (OQE) or the quadratic maximum likelihood estimator (QMLE); and was extensively studied by Hamilton (1997), Tegmark et al. (1997), Tegmark et al. (1998) and Seljak (1998).

The idea starts with a Gaussian likelihood function $L$ for measuring a data set $x$ for a given model that depends on parameters $\theta_\alpha$. We then pursue the most likely parameters $\hat{\theta}_\alpha$ for a fixed measurement $x$.

$$2L(x; \theta_\alpha) = \ln \det C + x^T C^{-1} x,$$

(1.7)

where $C = C(\theta_\alpha) \equiv \langle xx^T \rangle$ is the covariance matrix. The most likely parameters $\hat{\theta}$ can be found by maximizing this likelihood function: $L_{,\alpha}(\hat{\theta}) = 0$, where comma represents a partial derivative. In some simple cases, this formulation is enough to find an analytic expression. However, it is often solved numerically. Using an iterative Newton-Raphson method, we can further simplify what needs to be calculated.

$$\hat{\theta}_\alpha^{(X+1)} = \hat{\theta}_\alpha^{(X)} - \sum_{\alpha'} \langle L_{,\alpha'^\alpha} \rangle^{-1} \big|_{\hat{\theta}^{(X)}} \langle L_{,\alpha'^\alpha} \rangle \big|_{\hat{\theta}^{(X)}}^{-1},$$

(1.8)

where $X$ is the iteration number. Note that we compute the ensemble average of the second derivative (which is the Fisher matrix) instead of using the full curvature matrix as has been the convention. This is shown to converge faster while being unbiased by Taylor expanding a non-Gaussian likelihood (Madhavacheril et al., 2015). The covariance matrix

\(^3\)This approximation holds true for conserved sources, where the total mass along the line of sight is conserved under the transformation.
is the sum of signal and noise, \( C = S + N \); and in the specific case that is \( S = \sum_\alpha C_\alpha \theta_\alpha \), this simplifies to

\[
\hat{\theta}_\alpha^{(X+1)} = \sum_{\alpha'} \frac{1}{2} F_{\alpha\alpha'}^{-1} (d_{\alpha'} - b_{\alpha'}),
\tag{1.9}
\]

where

\[
d_{\alpha} = x^T C^{-1} C_\alpha C^{-1} x,
\tag{1.10}
\]

\[
b_{\alpha} = \text{Tr}(C^{-1} C_\alpha C^{-1} N),
\tag{1.11}
\]

and the estimated Fisher matrix is

\[
F_{\alpha\alpha'} \equiv \left\langle \frac{\partial^2 L}{\partial \theta_\alpha \partial \theta_{\alpha'}} \right\rangle = \frac{1}{2} \text{Tr}(C^{-1} C_\alpha C^{-1} C_{\alpha'}) .
\tag{1.12}
\]

The covariance matrices in the right hand side of equation 1.9 are computed using parameters from the previous iteration \( \theta_\alpha^{(X)} \).

### 1.3 The Intergalactic Medium

Primordial electrons and protons combine at \( z \sim 1100 \) to form neutral hydrogen. This epoch is known as the recombination, and probed by the cosmic microwave background (CMB) radiation experiments such as Cosmic Background Explorer (COBE) (Bennett et al., 1996), Wilkinson Microwave Anisotropy Probe (WMAP) (Bennett et al., 2013) and Planck (Planck Collaboration et al., 2020a). The universe then remains neutral until \( z \sim 10 \) when first galaxies form in halos with \( > 10^8 \text{M}_\odot \), and their highly energetic (\( > 13.6 \text{eV} \)) photons begin to reionize almost all hydrogen in the intergalactic medium (IGM). This violent process lasts until \( z \sim 6 \) and heats the gas from 10 K to \( 10^4 \text{K} \) (McQuinn, 2016). Figure 1.2 illustrates of this process.

Eventually, only a trace amount of neutral hydrogen survives at \( z < 5 \). Most are
clumped and shielded from the ionizing background in high-density systems; and 21-cm intensity mapping surveys such as CHIME (Bandura et al., 2014) and HIRAX (Newburgh et al., 2016) aim to observe the clustering of these at $z < 2$. Some reside diffusely in the IGM in equilibrium with the radiation. These diffuse neutral hydrogen clouds manifest themselves as absorption lines in quasar spectra by intervening the quasar emission at Lyman-alpha transition.

21-cm emission is the spontaneous transition of the electron-proton pair from the triplet state into the singlet state. These two states are modeled to be in equilibrium with respect to a spin-temperature $T_s$:

$$\frac{n_1}{n_0} = 3e^{-\hbar \nu_{10}/k_b T_s},$$

(1.13)

where $n_1$ is the number of atoms in the triplet state and $n_0$ is the number in the singlet state. $\nu_{10} = 1420.4$ MHz is the 21-cm emission line. Assuming $T_s \gg \hbar \nu_{10}/k_b$, 75% of neutral hydrogen are in the excited state, $n_1 = 3n_{\text{H}}/4$. The observed spectral intensity $I_{\nu}$ of this transition is measured by an effective brightness temperature $T_b$ in the Rayleigh-Jeans limit:
\[ I_\nu = 2k_B T_b \nu^2 / c^2 \] (Barkana & Loeb, 2007; Bull et al., 2015). The mean \( T_b \) of the 21-cm line at \( z \ll 2 \) can be estimated by the following expression (Chang et al., 2008):

\[ T_b = 0.3 \text{ mK} \left( \frac{\Omega_{\text{HI}}}{10^{-3}} \right) \left( \frac{\Omega_m + (1 + z)^3 \Omega_\Lambda}{0.29} \right)^{-1/2} \left( \frac{1 + z}{2.5} \right)^{1/2} \] (1.14)

Lastly, one studies the temperature fluctuations across the sky, \( \delta T_b \propto T_b \delta_{\text{HI}} \). Comprehensive formulations of 21-cm physics (with a focus on high \( z \)) can be found in Madau et al. (1997) and Furlanetto et al. (2006).

Theoretical foundations of the Ly \( \alpha \) forest were laid mostly in the 1990s (Bi et al., 1992; Hui & Gnedin, 1997; Gnedin & Hui, 1998; Weinberg et al., 1998; Croft et al., 1999b). Here, I will present a simple quantitative picture that ignores helium. The neutral hydrogen number density \( n_{\text{HI}} \) will be in ionization equilibrium with the recombination of protons and electrons:

\[ n_{\text{HI}} \Gamma_{\text{HI}} = n_p n_e \alpha_H(T), \] (1.15)

where \( \Gamma_{\text{HI}} \) is the ionization rate which depends on the ionization sources and history, and \( \alpha_H(T) = 4.2 \times 10^{-13} (T/10^4 \text{ K})^{-0.7} \) is the recombination coefficient that depends on temperature \( T \) (Hui & Gnedin, 1997). For \( z < 6 \), most of the neutral hydrogen is reionized, such that \( n_e \approx n_p \approx \rho_b / m_p \), where \( \rho_b \) is the baryonic mass density and \( m_p \) is the proton mass. We can find the number density of the neutral hydrogen from the equilibrium condition.

\[ n_{\text{HI}} = \frac{\alpha_H(T) \rho_b^2}{\Gamma_{\text{HI}} m_p^2} = \frac{\alpha_H(T) \rho_b^2}{\Gamma_{\text{HI}} m_p^2} (1 + \delta_b)^2 \] (1.16)

Furthermore, the density and temperature are connected via an equation of state \( T = T_0 (1 + \delta_b)^{\gamma - 1} \) (Hui & Gnedin, 1997). The observable is the transmitted flux fraction \( F = e^{-\tau} \), where \( \tau \) is the optical depth. For a photon emitted with frequency \( \nu_e \) at time \( t_e \)
and observed with $\nu_o$, the optical depth is given by

$$
\tau(\nu_o) = \int_{t_c}^{t_o} \left[ 1 - \exp \left( -\frac{h_P \nu'}{k_B T(t')} \right) \right] n_{\text{HI}}(t') \sigma_{\text{Ly}\alpha}(\nu') c dt',
$$

(1.17)

where $\nu'/1+z' = \nu_e/1+z_e = \text{and } \sigma_{\text{Ly}\alpha}(\nu')$ is the absorption cross-section. The exponential term can be ignored as IGM temperature $T \sim 10^4 \text{ K}$ and $T_{\text{Ly}\alpha} = h_P \nu_{\text{Ly}\alpha}/k_B = 1.2 \times 10^5 \text{ K}$.

Furthermore, we can assume absorption profile peaks at the Ly $\alpha$ transition and reach the following expression for the optical depth (Weinberg et al., 1998; Croft et al., 1998):

$$
\tau = 0.946 h(1+\delta_b)^{2.7-0.7\gamma} \left( \frac{1+z}{4} \right)^6 \left( \frac{\Omega_b h^2}{0.0125} \right)^2 \left( \frac{T_0}{10^4 \text{ K}} \right)^{-0.7} \left( \frac{\Gamma_{\text{HI}}}{10^{-12} \text{ s}^{-1}} \right)^{-1}
$$

(1.18)

### 1.4 Observations & Key Projects

A conventional redshift survey, such as The Sloan Digital Sky Survey (SDSS), combines redshift and angular position data and maps the three-dimensional matter distribution. As one of its great successes, SDSS clearly detected the acoustic peak due to the primordial photon-baryon coupling in the galaxy correlation function using over 45,000 luminous red galaxies at $z \approx 0.35$ in 2005 (Eisenstein et al., 2005). Improved measurements later achieved percent-level precision cosmology using a reconstruction technique (Eisenstein et al., 2007; Padmanabhan et al., 2012; Alam et al., 2017). The spectroscopic survey size will grow to 15 times the current SDSS sample as the upcoming Dark Energy Spectroscopic Instrument (DESI) will obtain spectra from about 30 million galaxies in five years (Levi et al., 2013; DESI Collaboration et al., 2016). As galaxy surveys grow in size and improve in analysis, novel techniques emerge and become competitive. In my thesis, two such techniques that probe neutral hydrogen are considered:

1. 21-cm intensity mapping surveys are being devised to measure the hyper-fine transition at low redshifts as a cheaper and faster complementary technique to the spectro-
scopic surveys. I studied a new reconstruction algorithm to improve their data.

2. The Lyman-alpha (Ly $\alpha$) forest is benefiting from the increase in sample size as more quasars are observed in spectroscopic surveys, and becoming the premier probe to measure the total neutrino masses and to study the IGM physics. I implemented and applied the optimal power spectrum estimator for the Ly $\alpha$ forest.

1.4.1 21-cm intensity mapping surveys

Instead of observing individual galaxies, a 21-cm intensity mapping survey measures the hyper-fine transition of neutral hydrogen in low resolution and observes galaxies in aggregate. This mapping technique has emerged as a promising probe for cosmological information on large-scales (Bull et al., 2015)—the 21-cm line is a natural redshift marker and we can focus on the large scales of interest without resolving each galaxy. This technique accurately traces the matter on cosmological scales (Chang et al., 2010; Switzer et al., 2013; Masui et al., 2013). Late-time 21-cm measurements such as CHIME (Bandura et al., 2014) and HIRAX (Newburgh et al., 2016) aim to observe large volumes of the universe in $0.8 < z < 2.5$ redshift range. However, foregrounds contaminate long-wavelength line-of-sight modes, and they are six orders of magnitude brighter than the signal.

The cosmic tidal reconstruction technique (Pen et al., 2012; Zhu et al., 2016; Zhu et al., 2018a) aims to recover these lost modes and to improve cross correlations of 21 cm surveys with CMB measurements and photo-z galaxy surveys (Furlanetto & Lidz, 2007; Adshead & Furlanetto, 2008; Masui et al., 2013). In my thesis, I investigated this reconstruction in redshift space with different low $k_{\parallel}$ cutoff foreground models. Furthermore, instrumental imperfections disseminate low $k_{\parallel}$ contamination into smaller-wavelength modes in both radial and angular directions (Morales et al., 2012). I considered this effect (called foreground wedge) in my thesis as well. Finally, I calculated a theoretical prediction for the efficiency
and observed behavior of this tidal reconstruction algorithm (Karaçaylı & Padmanabhan, 2019).

1.4.2 Lyman-alpha Forest

Matter can absorb as well as radiate. Neutral hydrogen gas between us and the source scatters photons as they redshift to the Ly$\alpha$ transition, manifesting itself as absorption lines called Ly$\alpha$ forest as can be seen in Fig. 1.3. The redshift of the quasar as well as any absorption line can easily be calculated from a spectrum:

$$z = \frac{\lambda - \lambda_{\text{Ly} \alpha}}{\lambda_{\text{Ly} \alpha}} - 1,$$

(1.19)

where $\lambda_{\text{Ly} \alpha} = 1216$ Å. A photon with emission (rest-frame) wavelength less than the Ly$\alpha$ wavelength, $\lambda_{\text{em}} < \lambda_{\text{Ly} \alpha}$, redshifts as it travels through the cosmic expansion. Once its wavelength stretches enough to be $\lambda_{\text{Ly} \alpha}$, it will be absorbed and scattered by the nearby neutral hydrogen gas, which forms the absorption lines in the spectrum. This absorption
spectrum maps the density fluctuations in the intergalactic medium (IGM) where deeper absorption lines correspond to higher density regions, so that the Ly$\alpha$ forest can be used to explore the large-scale structure. The Ly$\alpha$ forest technique can map out the matter distribution in vast volumes and at small scales ($r \lesssim 1\text{ Mpc}$) between $2 \lesssim z \lesssim 5$. At these redshifts, the structure formation is mildly non-linear, but the physics of the Ly$\alpha$ forest is further enriched by the thermal state and reionization history of the IGM (Hui & Gnedin, 1997; Gnedin & Hui, 1998). The line-of-sight flux power spectrum of the Ly$\alpha$ forest $P_{1D}$ has been at the frontier of constraining new physics including IGM thermal evolution (Boera et al., 2019; Walther et al., 2019), neutrino masses (Croft et al., 1999a; Seljak et al., 2006; Palanque-Delabrouille et al., 2015a,b; Yeche et al., 2017) and the nature of dark matter (Boyarisky et al., 2009; Viel et al., 2013; Baur et al., 2016; Iršič et al., 2017b; Garzilli et al., 2019).

Two categories of $P_{1D}$ data sets have emerged over the years. The first category contains thousands of low- to medium-resolution spectra obtained by the Extended Baryon Oscillation Spectroscopic Survey (eBOSS) (Dawson et al., 2016) and its predecessors; and the corresponding $P_{1D}$ estimates (McDonald et al., 2006; Palanque-Delabrouille et al., 2013; Chabanier et al., 2019). The upcoming Dark Energy Spectroscopic Instrument (DESI) (Levi et al., 2013; DESI Collaboration et al., 2016) aims to obtain approximately one million Ly$\alpha$ quasar spectra. With large sample size, this category can measure large scales to constrain cosmology and neutrino masses, but it is limited by noise and resolution at small scales. The second category contains tens to hundreds of high-resolution, high-S/N spectra obtained by various spectrographs including the High-Resolution Echelle Spectrograph (Vogt et al., 1994, HIRES), the Ultraviolet and Visual Echelle Spectrograph (Dekker et al., 2000, UVES) and X-Shooter spectrograph (Vernet et al., 2011). $P_{1D}$ estimates in this category often have been isolated to their respective data sets with 10–100 spectra (Viel et al., 2013; Walther et al., 2017; Iršič et al., 2017c; Boera et al., 2019; Day et al., 2019). The large-scale modes are inaccessible in this category due to lack of statistics, but these
spectra can probe extremely small scales which are crucial to constrain thermal state of the
IGM and some dark matter models.

Recent works utilize different methods in power spectrum estimation. For example, Palanque-Delabrouille et al. and Chabanier et al. applies fast Fourier transforms (FFT) in SDSS analyses, which require all pixels to be present, equally spaced and have uniform noise and resolution. These conditions are rarely met in real spectra due to masking of sky emission lines, high column density absorbers (HCD), bad pixels and sometimes metal contamination. As a result, Chabanier et al. apply up to 20% corrections to their power spectrum estimates. However, as the observational data improves in quality and quantity, precise measurements of one-dimensional Ly $\alpha$ power spectrum become important to constrain cosmological models with minimum error bars.

To this purpose, I implemented the quadratic maximum likelihood estimator (QMLE) or also known as the optimal quadratic estimator (OQE) (Hamilton, 1997; Tegmark et al., 1997, 1998; Seljak, 1998; McDonald et al., 2006). I proved its robustness against relevant problems including quasar continuum errors and gaps, showed that an input fiducial power spectrum improves the accuracy, and provided simple 5-yr forecasts for DESI $P_{1D}$ measurements (Karaçaylı et al., 2020). I also measured the highest-precision small-scale $P_{1D}$ with the largest high-resolution, high-S/N spectra using a combination of three public data sets (López et al., 2016; O’Meara et al., 2017; Murphy et al., 2019). With 538 quasars, this sample is seven times larger than Walther et al. (2017). Performing a crude, single-parameter Fisher forecast, I estimated that this result can improve warm dark matter mass uncertainties from $\sim 1.08$ keV to $0.45$ keV, which is a factor of 2.4 improvement.
Chapter 2

Tidal Reconstruction & 21-cm Intensity Maps

2.1 Introduction

21-cm intensity mapping has been a promising probe for cosmological information on large-scales (Bull et al., 2015). Late-time 21-cm measurements such as CHIME (Bandura et al., 2014) and HIRAX (Newburgh et al., 2016) aim to map large volumes of the universe through red-shifted neutral hydrogen emission line in $0.8 < z < 2.5$ redshift range. This method faces challenges from foregrounds that are six orders of magnitude brighter than the cosmological signal. Long-wavelength modes in the line-of-sight are contaminated by these spectrally smooth foregrounds. Furthermore, instrumental imperfections cause this noise to leak into high $k_\perp - k_\parallel$ modes and to form a foreground wedge in Fourier space (Morales et al., 2012). Since the long-wavelength radial modes play a significant role in cross correlations of 21-cm surveys with CMB measurements and photo-z galaxy surveys (Furlanetto & Lidz, 2007; Adshead & Furlanetto, 2008; Masui et al., 2013), recovering the lost modes is important to improve the cross correlation signal.

A procedure called cosmic tidal reconstruction has been developed to extract these lost modes from small-scale signal (Pen et al., 2012; Zhu et al., 2016, 2018b). Although
large-scale modes are not explicitly present, they leave an anisotropic imprint on small-scale modes through non-linear tidal interactions. To better visualize the nature of this interaction, we remind the reader the Moon’s tidal field on Earth. The gravitational potential of the Moon around the Earth’s centre can be Taylor expanded up to second order. The constant term can be ignored; the first derivative is the net gravitational force between two celestial objects. The second derivative of the potential causes deformations on Earth’s oceans and is the tidal field. Similarly, the cosmological gravitational potential can be separated into short- and long-wavelength parts. The latter can be expanded up to second order around an observer. This tidal field $t_{ij} \sim \Phi_{L,ij}$ then causes deformations in the local universe and introduces anisotropies to the local small-scale correlation function and power spectrum.

After finding the theoretical expression for the local small-scale power spectrum, we now need to estimate the underlying large-scale over-density field $\delta_L$ given an observation $\delta$. Assigning a Gaussian probability to measuring a data set $\delta$ for a given $\delta_L$, we can claim that our best estimate maximizes this probability. In technical terms, optimal quadratic estimators for $\delta_L$ are constructed assuming the observed density field $\delta$ is Gaussian. This assumption mandates mapping $\delta$ to a Gaussian field. Choosing robust $t_{ij}$ components is the last piece of the puzzle. Galaxies have peculiar velocities on top of the Hubble expansion. The line-of-sight positions are mismeasured because of these peculiar velocities. To minimize the so called redshift space distortions, two quadrupolar distortions of the tidal field in the plane perpendicular to the line-of-sight ($\gamma_1 = (\Phi_{L,xx} - \Phi_{L,yy})$ and $\gamma_2 = 2\Phi_{L,xy}$) are chosen. Even though both $\gamma_1$ and $\gamma_2$ are independent estimators of $\delta_L$ in theory, they can be combined into the three dimensional convergence field $\kappa_{3D}$ which is a better estimate for $\delta_L$ given that each $\gamma$ is zero in distinct $\hat{k}$ direction (Kaiser, 1992). A Wiener filter finally corrects for the bias and noise in $\kappa_{3D}$ itself.

A reconstruction algorithm first has to pass a test on ideal input by producing a highly correlated field. Then, it has to be robust against the obstacles arising from imperfections. Zhu et al. (2016) showed cosmic tidal reconstruction produced cross-correlations greater
than 0.9 on scales $k \lesssim 0.1 \, h/\text{Mpc}$ at $z = 0$ for full dark matter field in real space. In this work, we run the cosmic tidal reconstruction in real and redshift spaces to test the efficiency of the algorithm. We introduce the observational challenges to assess if the reconstruction is robust. In short, our work includes redshift space distortions, testing a range for low $k_{\parallel}$ data loss and including the foreground wedge.

The reader may be wondering how all these steps produce a correlated signal. The theoretical expectations are concealed behind tidal interactions, estimators and multiple Fourier transforms in implementation. Because of these complications, previous works relied only on simulations. In this work, we derive the analytic expressions for cross and power spectrum. We show that the cross correlations emerge from a modified bispectrum. This modification is due to Gaussian mapping of the initial over-density field. Our theoretical model can quantify the steep decline in reconstruction efficiency due to lost modes and the minor degradation due to peculiar velocities. It also promises a Wiener filter estimate when higher order terms are considered.

In Section 2.2, we review the tidal interaction theory and the construction of estimators. Section 2.3 describes the parameters for our N-body simulations and outlines the algorithm step by step. We present our results from these simulations in Section 2.4. We derive analytic cross and auto spectrum expressions in Section 2.5; and summarize in Section 2.6.

## 2.2 Theory

In this section, we first review the derivation of the local small-scale power spectrum (Zhu et al., 2016; Schmidt et al., 2014) and connect it to the construction of estimators (Lu & Pen, 2008). We refer the reader to the references for a longer discussion. Moreover, Akitsu et al. (2017) and Akitsu & Takada (2018) study the impact of the large-scale tidal field in a similar but different framework.
2.2.1 Tidal Interactions & Local Power Spectrum

The first ingredient of the cosmic tidal reconstruction is a theoretical prediction for local anisotropic features. What we call tidal interactions are the mechanism introducing these anisotropies. Zhu et al. has reviewed the traceless tidal field, and Schmidt et al. has studied the tidal interaction theory in detail using conformal Fermi Normal Coordinate frame (FNC). We find it sufficient to work in Newtonian picture for our discussion. Suppressing the time dependence, we start by decomposing the gravitational potential in Fourier space to separate it into short- and long-wavelength parts.

\[
\phi(x) = \int \frac{d^3k}{(2\pi)^3} \phi(k) e^{ik \cdot x} \quad (2.1)
\]

\[
= \int_{k<k_L} \frac{d^3k}{(2\pi)^3} \phi(k) e^{ik \cdot x} + \int_{k>k_L} \frac{d^3k}{(2\pi)^3} \phi(k) e^{ik \cdot x} \quad (2.2)
\]

\[
= \Phi_L(x, k_L) + \Phi_S(x, k_L) \quad (2.3)
\]

Now we consider a spherical volume of radius R around the origin such that \( k_L R \lesssim 1 \).

Suppressing \( k_L \) we can rewrite the long-wavelength potential as

\[
\Phi_L(x) = \int_{k<k_L} \frac{d^3k}{(2\pi)^3} \phi(k) e^{ik \cdot x} \quad (2.4)
\]

\[
= \int_{k<k_L} \frac{d^3k}{(2\pi)^3} \phi(k) \left( 1 + i k \cdot x - \frac{1}{2} k_i k_j x^i x^j \right) \quad (2.5)
\]

\[
= \Phi_L(0) + x \cdot \nabla \Phi_L(0) + \frac{1}{2} x^i x^j \Phi_{L,ij}(0). \quad (2.6)
\]

We can drop the constant term in the potential. The first derivative is a net force on the local universe; for large enough scales this can be ignored as well. The trace of \( \Phi_{L,ij}(0) \) describes if the local universe lives at an under- or over-dense region. The effect of this term is an overall change to the growth rate, which is not detectable and we ignore its
contribution here. So with a traceless tidal field \( t_{ij}(\tau) \equiv \Phi_{L,ij}(0) - \delta_{ij} \Phi_{L,kk}(0)/3 \), the gravitational potential \( \phi \) is given by

\[
\phi(\mathbf{x}) = \Phi_S + \epsilon_t \frac{1}{2} t_{ij}(\tau) x^i x^j,
\]

where small-scale gravitational potential \( \Phi_S \) obeys the Poisson equation

\[
\nabla^2 \Phi_S = \frac{3}{2} \Omega_m(\tau) \mathcal{H}^2(\tau) \delta_S.
\]

We will suppress subscript \( S \) to simplify our notation. We added \( \epsilon_t \) to keep track of the long-wavelength perturbations in powers of \( t_{ij} \). The tidal field can be written as \( t_{ij}(\tau) = T(\tau) t_{ij}^{(0)} \), where \( t_{ij}^{(0)} \) is the present value, \( T(\tau) = D(\tau)/a(\tau) \) and \( D(\tau) \) is the linear growth function.

The equation of motion for a particle in an expanding universe is

\[
\left[ \frac{d^2}{d\tau^2} + \mathcal{H} \frac{d}{d\tau} \right] \mathbf{x} \equiv \mathbb{D} \mathbf{x} = -\nabla_\mathbf{x} \phi,
\]

where we have defined the operator \( \mathbb{D} \). We can solve this equation using Lagrangian perturbation theory up to the nearest order in non-linear coupling between long- and short-wavelength perturbations.

\[
\mathbf{x} = \mathbf{q} + \Psi(\mathbf{q},\tau)
\]

\[
\Psi = \epsilon_s \Psi^{1s} + \epsilon_t \Psi^{1t} + \epsilon_s \epsilon_t \Psi^{1st} + \cdots
\]

The hypothetical particle in question still lives in the spherical volume, so \( x < R \) by construction. Additionally, \( 1s \) term is the linear solution and \( \epsilon_t \) signifies the contribution to the local small-scale fluctuations from the tidal field. We introduced \( \epsilon_s \) to keep track of the coupling order.
Let us first expand the over-density field using the mass conservation relation.

\[
\delta = \frac{1}{\det(\delta_i^j + \Psi_{i,j})} - 1 \quad \text{(2.12)}
\]

\[
= -\text{Tr}\Psi_{i,j} + \frac{1}{2}[(\text{Tr}\Psi_{i,j})^2 + \text{Tr}(\Psi_{i,j}^2)],
\quad \text{(2.13)}
\]

where \(\partial_j^i \equiv \Psi_{i,j}\). We set \(\Psi_{1t}^{i,i} = 0\) with hindsight because the tidal field is traceless \((t_{ii} = 0)\). The corresponding over-density in Lagrangian coordinates is

\[
\delta_{LG}(q) = \epsilon_s \delta_1^s + \epsilon_s \epsilon_t \left[\Psi_{i,j}^{1s} \Psi_{j,i}^{1t} - \Psi_{i,i}^{1st}\right].
\quad \text{(2.14)}
\]

We find the over-density field in Eulerian frame \(\delta_{EU}\) by demanding \(\delta_{LG}(q) = \delta_{EU}(x)\).

\[
\delta_{EU}(x) = \delta_{LG}(x - \Psi)
\]

\[
= \delta_{LG}(x) - \Psi \cdot \nabla \delta_{LG}(x) \quad \text{(2.15)}
\]

\[
\delta_{EU}(x) = \epsilon_s \delta_1^s(x) + \epsilon_s \epsilon_t \left[\Psi_{i,j}^{1s} \Psi_{j,i}^{1t} - \Psi_{i,i}^{1st}\right](x) - \epsilon_s \epsilon_t \Psi_{i}^{1t} \partial_i \delta_1^s(x) \quad \text{(2.16)}
\]

Then, the tidal field’s contribution to the local density fluctuations \(\delta_t = \delta_{EU} - \delta_1\) is

\[
\delta_t(x) = \left[\Psi_{i,j}^{1s} \Psi_{j,i}^{1t} - \Psi_{i,i}^{1st}\right](x) - \Psi_{i}^{1t} \partial_i \delta_1^s(x). \quad \text{(2.17)}
\]

Now, we go back to the equation of motion. We find the divergence using the chain rule

\[
\partial_i^\rho = \partial_i^x + \Psi_{j,i} \partial_j^x.
\]

\[
\square \Psi_{i,j} = -\nabla^2 \Phi - \Psi_{j,i} \left(\partial_i^x \partial_j^x \Phi + \epsilon_t t_{ij}\right) \quad \text{(2.19)}
\]

We can solve for the displacement field order by order. Here, we only summarize the
results.

\[ \Psi_{i}^{1s} = -D(\tau) \frac{\partial \delta_1}{\partial \tau} \delta_1^{(0)} \]
\[ \Psi_{t}^{1t} = -F(\tau) t_i^{(0)} q_i \]
\[ \Psi_{i,i}^{1st} = D_{1st}(\tau) D(\tau) t_i^{(0)} \frac{\partial q_i}{\partial \tau} \delta_1^{(0)} \]  \hspace{1cm} (2.20)

and the time evolution functions are given by

\[ F(\tau) = \int_{0}^{\tau} d\tau'' D(\tau'') \int_{\tau''}^{\tau} \frac{d\tau'}{a(\tau')} \]
\[ D_{1st}(\tau) = \int_{0}^{\tau} \frac{d\tau'}{H(\tau')} \frac{H(\tau')D(\tau') - D(\tau)H(\tau')T(\tau')}{D(\tau')} \]  \hspace{1cm} (2.21)

These functions in terms of \( z \) integral are in Appendix A.1.

Putting all these together and defining \( \alpha(\tau) \equiv -D_{1st}(\tau) + F(\tau) \), we find \( \delta_t \) in terms of \( \delta_1s \) and its spatial derivatives.

\[ \delta_t = t_i^{(0)} \left[ \frac{\alpha(\tau)}{\nabla^2} \delta_1^{(0)} + F(\tau) x_j \partial_i \right] \delta_1s(x, \tau). \]  \hspace{1cm} (2.22)

The local small-scale correlation function is defined with respect to a local origin, \( \xi(r) = \langle \delta(0) \delta(r) \rangle \) where \( \delta = \delta_1s + \delta_t \). Then, the nearest order power spectrum under tidal distortions is

\[ \tilde{P}_{1s}(k, \tau) = P_{1s}(k, \tau) + t_{ij}^{(0)} \dot{k}_i \dot{k}_j f(k, \tau) P_{1s}(k, \tau), \]  \hspace{1cm} (2.23)

where \( f(k, \tau) = -2\alpha(\tau) - F(\tau) d \ln P_{1s}(k, \tau)/d \ln k \). For clarity, we stress that 1s term is the linear solution and \( P_{1s} \equiv P_{l} \) is the linear power spectrum. Tilde represents tidal distortions, and \( \dot{k}_i = k_i/k \).

Analogous to weak lensing, we define \( \gamma_1 = (\Phi_{L,xx} - \Phi_{L,yy}) \) and \( \gamma_2 = 2\Phi_{L,xy} \) and
\[ \Delta P_{ls} = f(k, z)P_l(k, z) \left[ (\hat{k}_x^2 - \hat{k}_y^2)\gamma_1^{(0)} + 2\hat{k}_x\hat{k}_y\gamma_2^{(0)} \right]. \] (2.27)

Note that we ignore the contribution from the \( z \)-components of the tidal tensor to minimize the effects of redshift space distortions on our measurements.

### 2.2.2 Estimators

To estimate the underlying long-wavelength over-density field, we assign a Gaussian probability distribution for measuring a data set \( \tilde{\delta}(k) \) that depends on parameters \( \gamma_1 \) and \( \gamma_2 \) (Lu & Pen, 2008): 

\[ P(\tilde{\delta}(k); \gamma_1, \gamma_2) \propto |C|^{-1/2} \exp(-\tilde{\delta}^2/C^{-1})/2. \]

However, we pursue the best \( \gamma \) estimate of one data set. We define the likelihood function \( L \) as the negative logarithm of the probability distribution over parameters \( \gamma \) for a fixed \( \tilde{\delta}(k) \). We also assume a diagonal covariance matrix such that 

\[ C_{ij} = \langle \tilde{\delta}(k_i)\tilde{\delta}(k_j) \rangle = \delta_{ij}P_{tot}(k_i). \]

Then, the likelihood function in the continuum limit is

\[ L = \int d^3k \left[ \ln \tilde{P}_{tot}(k) + \frac{\tilde{\delta}(k)}{L^3\tilde{P}_{tot}(k)} \right], \] (2.28)

where an ensemble average gives \( \tilde{P}_{tot}(k) = \tilde{P}_{ls}(k) + P_N(k) \) and equation (2.26) provides the theoretical prediction for \( \tilde{P}_{ls}(k) \). The noise spectrum \( P_N(k) \) includes the effects of unmodelled non-linearities as well as instrumental noise.

We can construct the estimators by maximizing \( L \) with respect to \( \gamma_{1,2} \). Let us go through the calculation for \( \gamma_1 \).

\[ \frac{\partial L}{\partial \gamma_1} = \int \frac{d^3k}{(2\pi)^3} \left[ \frac{\tilde{P}_{tot}(k) - |\tilde{\delta}(k)|^2L^3}{\tilde{P}_{tot}^2(k)} \right] \frac{\partial \tilde{P}_{ls}}{\partial \gamma_1} \] (2.29)

\[ \frac{\partial \tilde{P}_{ls}}{\partial \gamma_1} = f(k, \tau)P_{ls}(k)(\hat{k}_1^2 - \hat{k}_2^2) \] (2.30)
Expanding the expression in parentheses to first order in estimates $\hat{\gamma}_{1,2}$ and limiting ourselves to quadratic estimators, we find

$$\hat{\gamma}_1 = \int \frac{d^3k}{(2\pi)^3} \frac{|\delta(k)|^2}{L^3} \frac{P_{1s}(k)}{P_{\text{tot}}^2(k)} f(k, \tau)(\hat{k}_1^2 - \hat{k}_2^2). \quad (2.31)$$

We implicitly absorbed the normalization coefficient into the Wiener filter. Furthermore, we can rewrite this expression by first defining

$$\tilde{\delta}_w^i(k) = \delta(k) \left[ \frac{P_{1s}(k)f(k)}{P_{\text{tot}}^2(k)} \right]^{1/2} i \hat{k}_i, \quad (2.32)$$

and then inverse Fourier transforming $\tilde{\delta}_w^i$. Integral over $k$ yields a Dirac delta function cancelling the other position variable.

$$\hat{\gamma}_1 = \int \frac{d^3x}{L^3} \left( \tilde{\delta}_{wx}(x) \tilde{\delta}_{wx}(x) - \tilde{\delta}_{wy}(x) \tilde{\delta}_{wy}(x) \right) \quad (2.33)$$

We have defined the tidal field as the second derivatives of the gravitational potential at an observer, so the tidal field and consequently $\hat{\gamma}_i$ are not functions of $x$. Nevertheless, the estimate $\hat{\gamma}_1$ is an average over local values as equation (2.33) indicates. Therefore, the estimate for $\gamma_1$ at a location $x$ is the expression in parentheses. These estimators are effectively weighted derivatives. Following the same steps, we find an expression for $\hat{\gamma}_2$.

$$\hat{\gamma}_1(x) = \delta_{wx}(x)\delta_{wx}(x) - \delta_{wy}(x)\delta_{wy}(x) \quad (2.34)$$

$$\hat{\gamma}_2(x) = \delta_{wx}(x)\delta_{wy}(x) + \delta_{wy}(x)\delta_{wx}(x) \quad (2.35)$$

Underlying density fluctuations and $\gamma_i$ are theoretically related through $k^2\gamma_1 \propto (k_x^2 - k_y^2)\delta_L/2$ and $k^2\gamma_2 \propto k_xk_y\delta_L$, so either $\gamma_i$ can be used to estimate $\delta_L$. Given $\gamma_1$ vanishes when $k_x = k_y$ and $\gamma_2$ vanishes when either $k_x$ or $k_y$ is zero, a better estimate for $\delta_L$ is the
three dimensional convergence field (Kaiser, 1992).

\[
\kappa_{3D}(k) = \frac{2k^2}{3(k_x^2 + k_y^2)^{3/2}} \left[ (k_x^2 - k_y^2)\hat{\gamma}_1(k) + 2k_x k_y \hat{\gamma}_2(k) \right] \tag{2.36}
\]

However, we still need to filter out noise and correct for multiplicative biases in \(\kappa_{3D}\).

The Wiener filter is constructed by minimizing the error \(e = \langle (W\kappa_{3D} - \delta)^2 \rangle\) with respect to \(W\), which yields

\[
W(k_\perp, k_\parallel) = \frac{\langle \kappa_{3D}\delta \rangle}{\langle \kappa_{3D}\kappa_{3D} \rangle}. \tag{2.37}
\]

We compute these expectation values by direct simulations, although we also present an analytic approach in Section 2.5.

### 2.3 Implementation

#### 2.3.1 Simulations

We run 10 simulations with \(1024^3\) dark matter particles using GADGET-2\(^1\) (Springel et al., 2001; Springel, 2005) in a box of side length \(L = 1.5 \text{ Gpc}/h\) with periodic boundary conditions. The boxes have the following cosmological parameters: \(\Omega_m = 0.276\), \(\Omega_b = 0.045\), \(\Omega_{\Lambda} = 0.724\), \(h = 0.7\), \(n_s = 0.961\) and \(\sigma_8 = 0.811\). The simulations start from \(z = 49\) using 2LPT initial conditions (Scoccimarro, 1998; Jenkins, 2010) constructed with the linear power spectrum from CAMB\(^2\).

In the following sections, we compare the reconstruction results at \(z = 0\) and \(z = 1\) in real and redshift space without any foreground subtraction. For every other case, we run the reconstruction only at \(z = 1\).

\(^1\)http://wwwmpa.mpa-garching.mpg.de/gadget/
\(^2\)https://lambda.gsfc.nasa.gov/toolbox/tb_camb_form.cfm
2.3.2 Method

The tidal reconstruction algorithm presented in Zhu et al. (2016) consists of two phases. The first phase computes $\kappa_{3D}$ from noisy over-density field, while the second phase applies a Wiener filter to $\kappa_{3D}$ to obtain the clean reconstructed field $\kappa$.

In the first phase:

1. We interpolate the density field to a grid and smooth it with a Gaussian window function.

2. We apply Gaussian mapping\(^3\) while keeping the standard deviation same ($\sigma(\delta_G) = \sigma(\delta_R)$) to obtain the Gaussianized over-density field $\delta_G(x)$ (Weinberg, 1992).

3. We construct $\delta_{wi}^G(x)$ using equation (2.32).

4. We estimate $\gamma$’s from equations (2.34) and (2.35).

5. We use equation (2.36) to determine the three dimensional convergence field $\kappa_{3D}$.

Note that 3 and 5 are done in Fourier space, while 4 is a configuration space operation. As is normal, all smoothing steps are done in Fourier space. We use FFTW\(^4\) for Fourier transforms on $1536^3$ grids and $R = 1.5$ Mpc/h as our smoothing radius, enough to suppress the interpolation kernel.

To estimate $\delta_{wi}^G$, we adopt $P_{ls}(k)$ as the linear power spectrum from CAMB and non-linear power spectrum for $\tilde{P}_{tot}(k)$. In redshift space we use a simple Kaiser form: $\tilde{P}^{(s)}_{tot}(k) = (1 + f\mu^2)^2 \tilde{P}_{tot}(k)$. The reconstruction works on $xy$ slices and does not have information in purely $z$ direction. As a result $\kappa_{3D}(k_\perp = 0, k_z)$ modes go to infinity and

\(^3\)Even though the log transform is a fast and simple Gaussianization procedure, the foreground subtraction models will cause $\delta < -1$ in our simulations, which hinders evaluating $\ln(1 + \delta)$. Also considering the possible numerical errors at void regions, we have chosen to map the field into a Gaussian distribution by preserving the ranking. Given that any interferometic map will have regions $\delta < -1$ with or without foreground subtraction, this mapping will be required in real data as well.

\(^4\)http://www.fftw.org
contain only noise. We set these modes to zero. Since peculiar velocities cause distortions
in $x_\parallel$ direction, we expect $\gamma_{1,2}$ not to be particularly affected by redshift space distortions.

$\kappa_{3D}$ is a biased estimate of $\kappa$; it is noisy and missing a normalization constant. We
construct Wiener filter from ten $\langle \kappa_{3D}\delta \rangle$ and $\langle \kappa_{3D}\kappa_{3D} \rangle$ pairs. To get better statistics, we bin
the spectra in $(k_\perp,k_\parallel)$. We linearly interpolate the Wiener filter in 2D using GSL\textsuperscript{5}. In the
second phase, the reconstructed field $\kappa$ is the filtered three dimensional convergence field.

$$\kappa(k) = \kappa_{3D}(k)W(k_\perp,k_\parallel)$$ (2.38)

We stress the ensemble (and angle) average in the Wiener filter expression. This filtering
does not perfectly recover the true over-density field.

## 2.4 Results

Throughout this work we plot the mean cross-correlation coefficients from N-body sim-
ulations between true over-density field $\delta$ and the reconstructed field $\kappa$, where
$$r_{\kappa\delta} = \frac{P_{\kappa\delta}}{\sqrt{P_\delta P_\kappa}}.$$ Error bars are computed from the standard deviation of ten simulation results.

### 2.4.1 Tests on the Reconstruction

We run the reconstruction on full dark matter field in real and redshift spaces at $z = 0$ and
$z = 1$. On the top panel of figure 2.1, solid lines represents the results for $z = 1$, whereas
dashed lines represents the results for $z = 0$. The results are similar for all cases, so we
explicitly show data points and error bars only for $z = 1$. The redshift space distortions
(red circles) degrade the correlation coefficient by less than 5%. Given the reconstruction
relies on non-linear coupling, its efficiency depends on redshift in a complicated way. For
example, purely linear regime ($z \to \infty$) has no coupling to exploit, whereas late times
\textsuperscript{5}https://www.gnu.org/software/gsl/
Figure 2.1: (Top) The cross-correlation coefficients $r_{\kappa\delta}(k)$ in real and redshift spaces without removing any modes at $z = 0$ and $z = 1$. The redshift space distortions cause $\approx 0.03$ decrease in the correlation coefficient. (Bottom) $r_{\kappa\delta}(k)$ after applying $k_\perp$ cutoff at redshift $z = 1$ in real space with $R = 1.5$ Mpc/h. Losing small-scale angular modes decreases the correlation coefficient.
Figure 2.2: The cross-correlation coefficients $r_{\kappa\delta}(k)$ at $z = 1$ with $R = 1.5 \text{ Mpc}/h$ in redshift space after applying the low $k_\parallel$ cutoff and removing foreground wedge. After decreasing approximately 0.15, the correlation coefficient is not sensitive to $k_\parallel^c$ up to 0.1 $h$/Mpc. Removing the wedge modes impairs the reconstruction and makes the $k_\parallel^c$ value irrelevant.
Figure 2.3: The cross-correlation coefficients $r(k_{\perp}, k_{\parallel})$ at $z = 1$ with $k_{\parallel}^c = 0.1\ h/\text{Mpc}$ in redshift space. (Top Left) A cartoon for foreground removal which does not include redshift space distortions. Removed wedge modes are represented by bright red colour. They also include high $k$. (Top Right) If only low $k_{\parallel}$ modes are removed, the reconstruction recovers correlations up to $k_{\perp} \approx 0.2\ h/\text{Mpc}$. As expected, low $k_{\perp}$ with $k_{\parallel} \gtrsim 0.02\ h/\text{Mpc}$ modes contain noise and are not correlated with the original over-density field. (Bottom) Removing wedge modes reduces $r(k_{\perp}, k_{\parallel})$ and limits the reconstruction to a triangle up to $k_{\perp} \approx 0.05\ h/\text{Mpc}$ with low $k_{\parallel}$.
has higher order coupling besides nearest order. Nevertheless, we do not see a strong improvement at higher redshift. Overall, the reconstruction keeps the correlations high ($\gtrsim 0.7$) up to $k \approx 0.1 \, h/\text{Mpc}$. However, our results are approximately 10\% worse than the results of Zhu et al. (2016). We attribute this difference to a combination of our lower particle density and higher smoothing radius.

The reconstruction is based on non-linear tidal interactions, so it needs small-scale modes to work efficiently. These modes are limited by the finite beam size, foreground wedge effect and the smoothing radius. Increasing the smoothing radius, we confirm the findings of Zhu et al.: The smoothing radius $R$ should not be larger than 5 Mpc/$h$. To simulate the finite beam, we remove high $k_\perp$ modes since they cannot be well resolved. We implement these as in the foreground case described below. Note that this cutoff makes small-scale signal weaker, but unlike Gaussian smoothing it is a real loss of information and not isotropic. The bottom panel of figure 2.1 demonstrates the reconstruction mostly
uses modes with $k_\perp \lesssim 1.0 \, h/\text{Mpc}$ and does not work well without modes with $k_\perp \gtrsim 0.3 \, h/\text{Mpc}$. Results in the redshift space are similar and are not depicted. In reality, how well an angular scale can be resolved depends on the baseline distribution for the experiment. A HIRAX-like experiment has access to a broad range of angular scales; $k_\perp < 1 \, h/\text{Mpc}$ are present, although with diminishing quantity (see White & Padmanabhan, 2017, figure 1). Moreover, the thermal noise power is a function of $k_\perp$. Simple estimates yield values between 150-600 Mpc$^3/h^3$ at $z = 1$ and $k_\perp = 0.2 \, h/\text{Mpc}$ (see White & Padmanabhan, 2017, Appendix A) and cross over at $k = 0.6 \, h/\text{Mpc}$ (Zhu et al., 2018b). Since our figures and results show simply removing modes, the reality will be between the cases we have presented.

### 2.4.2 Foregrounds

In this section, we investigate if the reconstruction can recover absent modes. Foregrounds from galactic and extra-galactic synchrotron and free-free emissions are spectrally smooth and contaminate modes with small $k_\parallel$. Without any angular resolution cutoff, we now turn to the astrophysical foregrounds problem at hand. These foregrounds have been implemented using a high-pass filter in Zhu et al. (2018b). However, estimator construction in Section 2.2.2 has motivated us to take a different approach. Instead of a high-pass filter, we adopt a sharp cutoff in $k_\parallel$ and assume $P_N(\mathbf{k})$ to be infinity for every mode set to zero. We first remove $k_z < k_\parallel^c$ modes from the original over-density field. Since these modes are purely noise, we set $\delta_g^{wi}(k_z < k_\parallel^c)$ in equation (2.32) to zero explicitly. The reconstruction recovers discarded modes in configuration space. We stress that these foregrounds cost high $k_\perp$ modes as well and reduce the reconstruction’s efficiency.

Instrumental limitations introduce another challenge by allowing low $k_\parallel$ foregrounds to leak into higher $k_\perp - k_\parallel$ modes (Morales et al., 2012). This effect forms a foreground wedge and contaminates modes with $|k_\parallel| \leq m(z)k_\perp$, where $m(z) = D_M(z)H(z)/(1 + z)$
and \( D_M(z) \) is the co-moving angular diameter distance (Seo & Hirata, 2016; White & Padmanabhan, 2017). We set these modes to zero to realize the foreground wedge. This effect costs even more small-scale modes and significantly limits the reconstruction’s efficiency.

We remind the reader of our two-phase procedure. The first phase uses the foreground subtracted over-density field. The second phase reverts to using the true over-density field \( \delta \) to construct the Wiener filter.

We test three values for \( k_c^c (0.02 \ h/\text{Mpc}, \ 0.05 \ h/\text{Mpc} \ \text{and} \ 0.1 \ h/\text{Mpc}) \) in real and redshift space at \( z = 1 \). The correlation coefficient falls sharply after the foreground removal since we remove high \( k_\perp \) modes in the process. However, we find the results are not truly sensitive to \( k_c^c \). The reconstruction can recover lost modes up to \( k_c^c = 0.1 \ h/\text{Mpc}. \) Figure 2.2 displays the results in redshift space. Real space results follow the same trend with a better performance similar to figure 2.1.

The reconstruction deteriorates more dramatically in the presence of a foreground wedge as seen in figure 2.2. We show the results for \( k_c^c = 0.1 \ h/\text{Mpc} \); changing \( k_c^c \) has no effect anymore since the additional information in low \( k_\perp \) triangle is insignificant.

In figure 2.3, we show the cross correlation coefficient in the anisotropic \( (k_\perp, k_\parallel) \) plane. The figure on the left is a cartoon for the foreground removal procedure, which does not depict redshift space distortions; and the removed wedge modes are represented in bright red. The middle panel shows the results for the low \( k_\parallel \) removal. Modes with \( k_\parallel < 0.1 \ h/\text{Mpc} \) are recovered up to \( k_\perp \approx 0.2 \ h/\text{Mpc} \), but the anisotropic nature of reconstruction causes a lower \( r(k) \) at \( k \approx 0.2 \ h/\text{Mpc} \) in figure 2.2. Without the wedge modes the reconstruction cannot recover \( k_\perp \gtrsim 0.1 \ h/\text{Mpc} \) modes, and its performance is limited around the wedge line with \( k_\perp \lesssim 0.05 \ h/\text{Mpc} \). The performance of the reconstruction on \( k_\parallel = 0 \) plane is shown in figure 2.4. The correlations are significantly higher on this plane as figure 2.3 signifies. These modes weigh more in weak lensing cross correlations.

The level to which the foreground wedge can be corrected will depend on the experiment
and is still unknown. So, we prefer isolating the wedge effect from low angular resolution, whereas in reality both effects will be present. The details of foregrounds and available angular modes will depend on the instrument and should be calculated into $P_N$ for an exact treatment. Although it would deteriorate the reconstruction further, combining these two does not necessarily negate the recovery, since there is an overlap between two cases.

### 2.5 Theoretical Expectations

The cosmic tidal reconstruction combines distinct theoretical pieces such as tidal interactions, maximum likelihood estimators and three dimensional convergence field, then adds another complication by requiring multiple Fourier transforms (see Section 2.3.2 for details). Due to these difficulties, previous works relied only on simulations to examine the reconstruction. Incorporating all the theoretical segments into one framework, we derive analytic expressions for cross and power spectrum. These expressions yield the correct structure in power spectrum and quantify the expectations in reconstruction efficiency due to lost modes and peculiar velocities. This framework also promises an estimate for Wiener filter when higher order terms are considered.

We start by defining the Gaussianized over-density field $\delta_G(x) = f_G[\delta_R(x); \sigma]$, where $\delta_R$ is smoothed with a Gaussian window function with radius $R$ and $f_G$ is the Gaussian mapping. Then, we rewrite equation (2.32) as $\delta_{G}^{wi}(k) = \delta_G(k)w(k)(ik_i/k)$. We transform equations (2.34) and (2.35) for $\gamma_i(x)$ into $k$ integrals by substituting Fourier transforms of $\delta_{G}^{usi}(x)$.

$$\gamma_i(x) = \int \frac{d^3kd^3q}{(2\pi)^6} e^{i(k+q)\cdot x} \tilde{T}_{\gamma_i}(k, q) \delta_{G}(k) \delta_{G}(q), \quad (2.39)$$
where

\[ \hat{T}_{\gamma_1}(k, q) = \left[ \frac{-k_x q_x - k_y q_y}{k q} \right] w(k) w(q) \]  

(2.40)

\[ \hat{T}_{\gamma_2}(k, q) = \left[ \frac{-k_x q_y + k_y q_x}{k q} \right] w(k) w(q). \]  

(2.41)

We warn the reader that \( q \) in this context does not represent Lagrangian coordinates.

Note that \( w(k) = w(-k) \) and therefore \( T_{\gamma_i} \) is symmetric and has the following properties:

\[ \hat{T}_{\gamma_i}(k, k') = \hat{T}_{\gamma_i}(k', k) = \hat{T}_{\gamma_i}(-k, -k') \]  

(2.42)

\[ = -\hat{T}_{\gamma_i}(k, -k') = -\hat{T}_{\gamma_i}(-k, k'). \]  

(2.43)

Now, we find the Fourier transforms of \( \gamma \)'s.

\[ \gamma_i(k) = \int \frac{d^3x}{(2\pi)^3} e^{-ik \cdot x} \gamma_i(x) \]  

(2.44)

\[ = \int \frac{d^3q d^3p}{(2\pi)^6} e^{i(q+p-k) \cdot x} \hat{T}_{\gamma_i}(q, p) \delta_G(q) \delta_G(p). \]  

(2.45)

The integral over \( x \) yields a Dirac delta function which integrates out \( p \). Consequently, the local estimates for each \( \gamma \) becomes

\[ \gamma_i(k) = \int \frac{d^3q}{(2\pi)^3} \hat{T}_{\gamma_i}(q, k - q) \delta_G(q) \delta_G(k - q) \]  

(2.46)

in Fourier space. Equation (2.46) easily extends to an expression for the three dimensional convergence field \( \kappa_{3D}(k) \). We prefer defining \( p \equiv k - q \) to further simplify the notation.

Using equation (2.36), we arrive at

\[ \kappa_{3D}(k) = \int \frac{d^3q}{(2\pi)^3} \hat{T}_{\kappa_{3D}}(q, p) \delta_G(q) \delta_G(p), \]  

(2.47)
where \( \tilde{T}_{\kappa_3D}(q, p) \) is given by equation (2.36) with \( \tilde{T}_{\gamma_i} \) instead of \( \gamma_i \). The exact form of \( \tilde{T}_{\kappa_3D} \) can be found in Appendix A.2.

Our conventions for the cross spectrum and the power spectrum are the same:

\[
\langle \kappa_{3D}(k) \delta(k') \rangle = (2\pi)^3 \delta_D(k + k') P_{\kappa_3D\delta}(k). \tag{2.48}
\]

Having a direct link between \( \delta \) and \( \kappa_{3D} \), we are now able to find these spectra. The cross spectrum is

\[
P_{\kappa_3D\delta}(k) = \int \frac{d^3 q}{(2\pi)^3} B_T(q, p) \tilde{T}_{\kappa_3D}(q, p), \tag{2.49}
\]

where we have defined a modified bispectrum \( B_T \) as

\[
\langle \delta_G(q) \delta_G(k - q) \delta(k') \rangle = (2\pi)^3 \delta_D(k + k') B_T(q, p). \tag{2.50}
\]

Although assuming \( \delta_G \) as the linear field \( \delta_l \) is tempting, this assumption results in a very small bispectrum in low \( k \), since \( \delta \) also approximates the linear field at these scales. Given \( \delta_l \) is Gaussian and has zero bispectrum, \( B_T \propto \langle \delta_l^3 \rangle = 0 \). From this short discussion, we conclude that second order corrections in \( \delta_G \) have to be taken into account. This makes sense considering tidal reconstruction makes use of mode coupling terms which do not exist in linear theory. To find these necessary higher order contributions, we need to take a closer look at the Gaussianization procedure. Since the logarithmic transform is easier to handle (Neyrinck et al., 2011; McCullagh et al., 2016), whereas the exact Gaussian mapping is analytically complicated, we assume

\[
\delta_G(x) = \ln(1 + \delta_R(x)) - \langle \ln(1 + \delta_R(x)) \rangle \tag{2.51}
\]

\[
\approx \delta_R - \frac{\delta_R^2}{2} + \frac{\delta_R^3}{3} - \frac{\delta_R^4}{4} + \cdots - \text{const}. \tag{2.52}
\]

The constant average term is only in \( k = 0 \) mode and can be ignored. We ignore smoothing
for simplicity, and refer the reader to Appendix A.3 for expressions with smoothing. Taking the Fourier transform and using second-order perturbation results (Jain & Bertschinger, 1994), we arrive at

$$\delta_G(k) \approx \delta_l(k) + \int \frac{d^3q}{(2\pi)^3} F_{2,G}(q,p)\delta_l(q)\delta_l(p),$$

(2.53)

where $F_{2,G} = F_2 - \frac{1}{2}$ and

$$F_2(k_1, k_2) = \frac{5}{7} + \frac{k_1 \cdot k_2}{2} \left[ \frac{1}{k_1^2} + \frac{1}{k_2^2} \right] + \frac{2}{7} \frac{(k_1 \cdot k_2)^2}{k_1^2 k_2^2}.$$

(2.54)

Having these expressions, we can now calculate the modified bispectrum $B_T$. Its only difference from normal bispectrum $B$ calculation is $-1/2$ term in $F_{2,G}$, so this step is not too cumbersome. Then,

$$B_T(q, p) = B(q, p, -k) - P_l(k)(P_l(p) + P_l(q)),$$

(2.55)

where $P_l$ is the linear power spectrum and

$$B(k_1, k_2, k_3) = 2F_2(k_1, k_2)P_l(k_1)P_l(k_2) + \text{cyc.}$$

(2.56)

A similar calculation shows that the four-point function gives the power spectrum of $\kappa_{3D}$. We remind the reader that the four-point function $\langle \delta^4 \rangle$ can be decomposed into three cyclic $\langle \delta^2 \rangle^2$ terms and a connected term. The trispectrum is the Fourier transform of the connected four-point correlation function. In this estimation, assuming $\delta_G$ to be the linear field yields a zero trispectrum, but a non-zero four-point function with one vanishing cyclic term.

$$P_{\kappa_{3D}}(k) = 2 \int \frac{d^3q}{(2\pi)^3} F_{\kappa_{3D}}^2(q,p)P_l(q)P_l(p)$$

(2.57)
As we have discussed in the cross spectrum case, $\delta_G \approx \delta_l$ produces uncorrelated $\kappa_{3D}$.
Hence, we would expect this equation to underestimate the true power.

With all limitations in mind, we can build a theoretical Wiener filter using equations (2.49) and (2.57), and find the reconstructed power spectrum $P_\kappa$.

$$P_\kappa(k) = \frac{P^2_{\kappa_{3D}\delta}}{P_{\kappa_{3D}}}. \quad (2.58)$$

### 2.5.1 Results

Having performed extensive numerical analyses in Section 2.4, we now test our theoretical framework to understand the emergent features and limited efficiency of cosmic tidal reconstruction. Our focus is on equations (2.49) and (2.57). We include smoothing in our calculations (see Appendix A.3) and take $R = 1.5 \text{ Mpc}/h$ as before.

The volume integration is simpler in spherical coordinates.

$$\int d^3q \rightarrow \int_{q_i}^{q_f} q^2 dq \int_{-1}^{1} dv \int_{-\pi}^{\pi} d\phi, \quad (2.59)$$

where $\theta$ is the azimuthal angle with respect to the line-of-sight and $v = \cos \theta$. We apply two more transformations. First, we write the $\phi$ integral in terms of $t = \cos \phi$, since this integrand depends only on $\cos \phi$ and it is symmetric under $\phi \rightarrow -\phi$. Second, we use logarithmic spacing in $q$ integration. Putting these together, the volume integration becomes

$$\int d^3q \rightarrow \int_{\ln q_i}^{\ln q_f} q^3 \ln q \int_{-1}^{1} dv \int_{-\pi}^{\pi} \frac{2dt}{\sqrt{1-t^2}}. \quad (2.60)$$

We use GSL’s Monte-Carlo integration library to compute these integrals. Assuming the same cosmology and CAMB power spectra as before, we integrate from $q_i = 10^{-4} h/\text{Mpc}$ to $q_f = 2 h/\text{Mpc}$ at $z = 0$. We form an equally spaced, 50x50 grid for $(k_\perp, k_z)$
Figure 2.5: Power spectrum $P_\kappa$ of one simulation run (Top) compared to the theoretical estimate (Bottom). Even though theory correctly predicts the structure in the power spectrum, it yields larger $P_\kappa$ than numerical simulations. Our $P_{\kappa, 3D}$ estimate is actually uncorrelated due to nearest order approximation, and lacks the reconstruction signal. This lack of power in $\kappa_{3D}$ results in high $P_\kappa$. 
with values between 0.001 $h$/Mpc and 0.5 $h$/Mpc including both ends.

Figure 2.5 shows the theoretical $P_\kappa(k_\perp, k_z)$ on the bottom panel and a representative simulation on the top panel. While their structures and shapes agree, the theoretical $\kappa$ power spectrum is off by one order of magnitude. As we have discussed in the previous section, our theoretical $P_{3D}$ estimate lacks significant power from reconstructed field because of the lowest order expansion. Since $P_\kappa = P^2_{3D\delta}/P_{3D}$, the lack of power in $\kappa_{3D}$ results in high theoretical $P_\kappa$.

After confirming our theoretical framework in anisotropic power spectrum, we assess which modes contribute the most to the reconstructed field $\kappa$. We start by rewriting the cross spectrum.

$$P_{3D\delta} = \int_{q_i}^{\ln q_f} q' d\ln q' \int_{-1}^{1} dv' I(q', v')$$

(2.61)

Because $I(q, v)$ is not symmetric with respect to $v$ (i.e. $I(q, v) \neq I(q, -v)$), we define $\bar{I}(q, v) = I(q, v) + I(q, -v)$ such that $v \in [0, 1]$. To quantify the importance of modes, we integrate the cross spectrum from $(q_i, 0^\circ)$ to $(q, \theta)$, then divide this integral by the true value.

$$\alpha(q, \theta) = \frac{1}{P_{3D\delta}(k_\perp, k_z)} \int_{q_i}^{\ln q} q' d\ln q' \int_{v}^{1} dv' \bar{I}(q', v')$$

(2.62)

We evaluate the ratio $\alpha(q, \theta)$ at a fixed $(k_\perp = 0.05h$/Mpc, $k_z = 0.02h$/Mpc) value at $z = 0$ on a grid equally sampled with 100 $q$ points in $(10^{-4}, 1.5]$ $h$/Mpc and 90 equally spaced $\theta$ points in $(0, 90)$. $P_{3D\delta}$ is evaluated with $q_f = 2.5$ $h$/Mpc.

The value of a point $(q, \theta)$ in figure 2.6 represents the fraction of cross correlation signal that can be achieved using modes up to $q$ and angle $\theta$ with respect to the line-of-sight. The reconstruction prefers $q_\perp$ modes and reaches its peak efficiency at $q \approx 1.2$ $h$/Mpc when modes up to $\theta = \pi/2$ are accessible as the smallest fluctuations are suppressed by Gaussian smoothing. We added the wedge line for $z = 1$ for reference. Foreground wedge impedes access to higher $\theta$ modes and restricts the efficiency of reconstruction to approximately 0.2. Redshift space distortions affect the reconstruction in reverse direction, contaminating low
Figure 2.6: The fraction of cross correlation signal $\alpha(q, \theta)$ that can be achieved using modes up to $q$ and angle $\theta$ with respect to the line-of-sight at $z = 0$. The reconstruction prefers $q_\perp$ modes and reaches its peak efficiency approximately at $(1.2 \, h/\text{Mpc}, \pi/2)$. Foreground wedge would restrict the reconstruction efficiency to approximately 0.2.

$\theta$ modes. These are not significant in the reconstruction, explaining the weak dependence in figure 2.1.

### 2.6 Summary

21-cm intensity mapping surveys lose large-scale radial signal to astrophysical smooth foregrounds. These foregrounds then contaminate smaller-scale radial and angular modes. Recovering the lost data (low $k_\parallel$ modes) will improve BAO measurements and cross correlations of 21-cm signal with CMB measurements and photo-z galaxy surveys.

The large-scale over-density field deforms the local universe around an observer by a non-zero second derivative of the long-wavelength gravitational potential. This deformation
imprints anisotropic features in the local small-scale power spectrum and allows us to reconstruct the lost data.

The second derivative of $\Phi_L$ gives the tidal field $t_{ij}(\tau) \equiv \Phi_{L,ij}(0) - \delta_{ij} \Phi_{L, kk}(0)/3$. We choose two components ($\gamma_1$ and $\gamma_2$) to minimize redshift space distortions on the reconstruction. The estimators for each $\gamma$ can be constructed assuming the observed $\delta$ is a Gaussian field. In return, this assumption requires us to Gaussianize the observed $\delta$. The three dimensional convergence field $\kappa_{3D}$ is a linear combination of $\gamma_1$ and $\gamma_2$; and it is a better estimate for the underlying large-scale over-density field. We correct $\kappa_{3D}$ for bias and noise with a Wiener filter constructed from ten N-body simulations.

In this work, we reviewed non-linear tidal interactions using Lagrangian perturbation theory and explored the efficiency of reconstructing long-wavelength modes from local small-scale fluctuations. We performed new tests on cosmic tidal reconstruction by adding redshift space distortions, investigating a range of $k_\parallel$ and $k_\perp$ data loss and foreground wedge data loss. We also presented a novel theoretical framework to study the reconstruction, which can predict its efficiency and produce theoretical estimates for Wiener filter if improved.

We found the cross-correlation coefficient $r$ between the reconstructed field and true over-density field is above 0.7 until $k = 0.1 \, h/\text{Mpc}$ in both spaces at two redshifts. Our numerical tests also confirmed the reconstruction is robust against peculiar velocities with minor degradation (less than 5%) in $r$. This validates choosing quadrupolar distortions in the plane perpendicular to the line-of-sight, $\gamma_1$ and $\gamma_2$.

To assess 21-cm intensity mapping challenges, we tested the reconstruction against astrophysical foregrounds. We modelled spectrally smooth foregrounds as a sharp cutoff below $k_\parallel^c$ and found that the cross-correlation coefficient decreases by 20%, but it is not sensitive to the cutoff value when $k_\parallel^c < 0.1 \, h/\text{Mpc}$. However, tidal reconstruction does not recover modes isotropically and specifically does not recover low $k_\perp$-high $k_\parallel$ modes. Given this feature, the reconstruction recovers modes $k_\parallel < 0.1 \, h/\text{Mpc}$ up to
$k_\perp = 0.2 \, h/{\text{Mpc}}$. Moreover, smooth foregrounds will leak into higher $k_\perp - k_\parallel$ modes due to imperfect instrumentation. We modelled this foreground wedge by removing modes obeying $|k_\parallel| \leq k_\perp D_M(z)H(z)/(1 + z)$. After these modes are subtracted, $k_\parallel^2$ becomes irrelevant and the reconstruction deteriorates to $r(k < 0.1 h/{\text{Mpc}}) \lesssim 0.5$. In the anisotropic picture, modes around the wedge slope can be recovered with $r \sim 0.5$. On $k_\parallel = 0$ plane, the wedge case performance shows mild improvement, whereas other cases do significantly better.

Finally, we incorporated all the theoretical segments into one framework, and showed that the cross correlations arise from a modified bispectrum and the power spectrum of $\kappa_{3D}$ originates from the four-point function. Our framework revealed a similar structure in $P_\kappa$, but missed the reconstructed signal due to lowest order expansion in the four-point function. Using the same framework, we estimated what modes are most important for efficient reconstruction. Our theoretical investigation predicted performance variations: steep decline when wedge modes are lost and minor degradation in redshift space. We expect one-loop corrections will decrease the discrepancy between simulations and theoretical predictions; it could construct a capable Wiener filter as well.

We have investigated the angular resolution and foregrounds individually. The available modes will depend on the details of the instrument. An exact treatment should be calculated into $P_N$ for real data applications. Since the modes above the wedge are less valuable, this should still produce correlations $r \sim 0.4$.

Even though combining low angular resolution and the foreground wedge will not erase the correlations completely, we find the overall performance not as high as we would like. These challenges hint at utilizing the distortions in the radial direction. The trade-off for using all $t_{ij}$ components is to extend the study to redshift space. Since small scales matter, the instrument design should take into account of resolving smaller scales and mitigating foreground wedge as much as possible as well. We hope that our work stresses the necessity and importance of such further investigations.
Throughout our work we assumed neutral hydrogen perfectly traces dark matter. The neutral hydrogen bias should be modelled for further investigation. A simple proposal would be to take haloes as tracers.
Chapter 3

Optimal 1D Lyman-alpha Forest Power Spectrum Estimator

3.1 Introduction

Through absorption lines in quasar spectra, the Ly-α forest technique can probe matter in vast volumes far into the past and at smaller scales than galaxy surveys, that are shaped by the thermal state of the gas and reionization history of the universe (Hui & Gnedin, 1997; Gnedin & Hui, 1998). Connecting flux fluctuations in quasar spectra to physical parameters relies on multiple demanding steps in a typical analysis. The first step is to summarize the statistical information contained in millions of pixels across thousands of spectra using the correlation function or the power spectrum. Second, one relies on numerical simulations to relate the matter fluctuations to the neutral hydrogen which reionizes until $z \sim 6$ to obtain mock quasar spectra. The physical parameters are then mapped to the statistics using large numbers of these mocks with different parameters. The final step constrains physical parameters by performing a likelihood analysis on the observed statistics using this mapping and a prior.

The Ly-α forest technique already proved to be fruitful. The Extended Baryon Oscillation Spectroscopic Survey (eBOSS) (Dawson et al., 2016) and its predecessors successfully
measured baryon acoustic oscillations in large-scale 3D correlations of the Ly-\(\alpha\) forest (Slosar et al., 2011, 2013; Busca et al., 2013; Font-Ribera et al., 2014; Delubac et al., 2015; du Mas des Bourboux et al., 2017; Bautista, Julian E. et al., 2017; Blomqvist et al., 2019; de Sainte Agathe et al., 2019). It has also emerged as a promising tool to investigate intergalactic medium (IGM) thermal evolution (Boera et al., 2019; Walther et al., 2019), to constrain neutrino masses (Croft et al., 1999a; Seljak et al., 2006; Palanque-Delabrouille et al., 2015a,b; Yeche et al., 2017) and to probe the nature of dark matter (Boyarsky et al., 2009; Viel et al., 2013; Baur et al., 2016; Iršič et al., 2017b; Garzilli et al., 2019).

The line-of-sight flux power spectrum of the Ly-\(\alpha\) forest has been at the frontier in new physics by being sensitive to medium to small scales. McDonald et al. (2006), Palanque-Delabrouille et al. (2013) and Chabanier et al. (2019) measured this 1D power spectrum at large to medium scales (\(0.001 \text{ s} \text{ km}^{-1} \leq k \leq 0.02 \text{ s} \text{ km}^{-1}\)) using thousands of quasar spectra, whereas Viel et al. (2013), Iršič et al. (2017c), Walther et al. (2017), Boera et al. (2019) and Day et al. (2019) pushed the measurement to smaller scales (\(k \leq 0.1 \text{ s} \text{ km}^{-1}\)) using few but high-resolution spectra.

These recent works utilize three different methods in power spectrum estimation. Palanque-Delabrouille et al. and Chabanier et al. apply fast Fourier transforms (FFT), which require all pixels to be present, equally spaced and have uniform noise and resolution. These conditions are rarely met in real spectra due to masking of sky emission lines, high column density absorbers (HCD), bad pixels and sometimes metal contamination. As a result, Chabanier et al. apply up to 20% corrections to their power spectrum estimates. In their small-scale measurement, Walther et al. and Day et al. use Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) as this method allows for masking. However, neither method is able to weight pixels with respective noise estimates and cannot account for the time evolution within a spectrum without splitting the spectrum into multiple chunks; and both are limited by S/N and resolution in their quasar samples. Third method is the likelihood maximization. Palanque-Delabrouille et al. also implemented a direct maximization of the
likelihood function, which they found sensitive to the implementation details and noise in spectra. Our focus in this work is the faster and more stable quadratic maximum likelihood estimator (QMLE), which McDonald et al. applied while splitting the spectra into two chunks because of computational limitations. However, with recent developments in computing, QMLE promises its full strength by careful application to 1D power spectrum.

Our work focuses on developing and applying an improved QMLE to measure the 1D power spectrum. We meticulously apply the formalism to maximize the information extraction. Our method readily handles the redshift evolution by interpolating pixel pairs to two redshift bins. This enables us to keep the full forest and all pixel pairs. We also introduce a baseline estimate that improves the accuracy by alleviating discrete band powers. As the Dark Energy Spectroscopic Instrument (DESI) (Levi et al., 2013; DESI Collaboration et al., 2016) comes online to find abundance of quasar spectra, our improved QMLE can exploit the most information and make the best measurement.

This chapter is organized under six sections. We summarize relevant general and specific formula for QMLE in section 3.2. This section also briefly discusses a continuum limit to clarify what QMLE actually yields. Section 3.3 outlines the algorithm and further provides details on our implementation, including validation with synthetic spectra. We then move to applying our method to simulated simple DESI data in section 3.4. In this section, we examine the effects of gaps and continuum errors, the advantage of fiducial power, Fisher matrix approximation and five-year forecasts for DESI with different spectral qualities. Section 3.5 discusses the finer details of our method, such as interpretation of QMLE results, its advantages and possible problems. Finally, we summarize this work in section 3.6.
3.2 Method

The optimal power spectrum estimator was extensively studied by Hamilton (1997), Tegmark et al. (1997), Tegmark et al. (1998) and Seljak (1998). We first assign a Gaussian probability distribution that depends on parameters $\theta_\alpha$ for measuring a data set $x$. We define the likelihood function $L$ as twice the logarithm of this probability and pursue the most likely parameters $\hat{\theta}_\alpha$ for a fixed $x$.

$$2L(x; \theta_\alpha) = \ln \det C + x^T C^{-1} x,$$  \hspace{1cm} (3.1)

where $C = C(\theta_\alpha) \equiv \langle xx^T \rangle$. The most likely parameters $\hat{\theta}$ can be found by maximizing this likelihood function: $L,\alpha(\hat{\theta}) = 0$, where comma represents a partial derivative, which can be iteratively solved using the Newton-Raphson method:

$$\hat{\theta}_\alpha^{(X+1)} = \hat{\theta}_\alpha^{(X)} - \sum_{\alpha'} \langle L,\alpha\alpha' \rangle^{-1} \big|_{\hat{\theta}^{(X)}} L,\alpha' \big( \hat{\theta}^{(X)} \big), \hspace{1cm} (3.2)$$

where $X$ is the iteration number. Note that we compute the ensemble average of the second derivative (which is the Fisher matrix) instead of using the full curvature matrix as has been the convention.

Our goal is to estimate the power spectrum of the observed spectra, so we take $\theta_\alpha$ to be the power spectrum estimates. Furthermore, we would like to estimate deviations from a fiducial power spectrum such that $P(k, z) = P_{\text{fid}}(k, z) + \sum_\alpha w_\alpha(k, z) \theta_\alpha$, where $w_\alpha(k, z)$ are the functional forms for deviations and $\theta_\alpha$ are the amplitudes (Font-Ribera et al., 2018). Then, the covariance matrix is the sum of signal and noise as usual, $C = S + N$; and by extension $S = S_{\text{fid}} + \sum_\alpha Q_\alpha \theta_\alpha$, where $Q_\alpha = \partial C / \partial \theta_\alpha$. The fiducial power needs to be subtracted from the estimate. We denote its contribution by $t_\alpha$ below, which can be immediately calculated by substituting $N \to N + S_{\text{fid}}$ to the quadratic estimator equation.
in the references.

\[ \hat{\theta}_{\alpha}^{(X+1)} = \sum_{\alpha'} \frac{1}{2} F_{\alpha \alpha'}^{-1} (d_{\alpha'} - b_{\alpha'} - t_{\alpha'}), \]  
(3.3)

where

\[ d_{\alpha} = x^T C^{-1} Q_{\alpha} C^{-1} x, \]  
(3.4)

\[ b_{\alpha} = \text{Tr}(C^{-1} Q_{\alpha} C^{-1} N), \]  
(3.5)

\[ t_{\alpha} = \text{Tr}(C^{-1} Q_{\alpha} C^{-1} S_{\text{fid}}), \]  
(3.6)

and the estimated Fisher matrix is

\[ F_{\alpha \alpha'} \equiv \left\langle \frac{\partial^2 L}{\partial \theta_\alpha \partial \theta_{\alpha'}} \right\rangle = \frac{1}{2} \text{Tr}(C^{-1} Q_{\alpha} C^{-1} Q_{\alpha'}). \]  
(3.7)

The covariance matrices in the right hand side of equation 3.3 are computed using parameters from the previous iteration \( \theta_{\alpha}^{(X)} \).

3.2.1 Ly \( \alpha \) Forest Specifics

In the Ly-\( \alpha \) forest analysis, our data set \( x \) is a collection of pixels representing the normalized flux fluctuations \( \delta_F \). Assuming different quasar spectra are uncorrelated, the covariance matrix becomes block diagonal, where only the correlations within a spectrum are non-zero\(^1\). For example, stacking three quasar spectra \( \delta_{1,2,3} \) yields:

\[ \delta_F = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix}, \quad \rightarrow C = \begin{pmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & C_3 \end{pmatrix}. \]  
(3.8)

\(^1\)The 3D analysis inherently needs these correlations between spectra, so this would not hold true.
This block diagonal structure simplifies equation (3.3) as well: The Fisher matrix $F_{\alpha\alpha'}$ and the expression in parentheses can be computed for each quasar, then accumulated, i.e. $F = \sum_q F_q$ etc.

The correlation (signal) between pixels depends on their velocity separation, underlying power spectrum and spectrograph window function. We convert a pixel’s wavelength to velocity using logarithmic spacing.

$$v_i = c \ln(\lambda_i/\bar{\lambda})$$  \hspace{1cm} (3.9)

$$z_i = (1 + \bar{z})e^{v_i/c} - 1,$$  \hspace{1cm} (3.10)

where $\bar{\lambda}$ and $\bar{z}$ are the median wavelength and the median redshift of the spectrum respectively\(^2\). In general, the signal is multiplied with a resolution matrix $R$, such that $\tilde{s} = Rs$ and therefore $\tilde{S} = RSR^T$ (Bolton & Schlegel, 2010). DESI will provide this resolution matrix in its pipeline, which will be one of QMLE’s strengths for future analyses. For the rest of the chapter, we make the approximation that the resolution does not change with wavelength. Then, the signal becomes the correlation function convolved with the spectrograph resolution, which is the power spectrum multiplied with the spectrograph window function $W(k)$ in Fourier space.

$$S_{ij}^{\text{fid}} = \int_0^\infty \frac{dk}{\pi} \cos(kv_{ij})W^2(k)P_{\text{fid}}(k, z_{ij}),$$  \hspace{1cm} (3.11)

where $v_{ij} \equiv v_i - v_j$ and $1 + z_{ij} \equiv \sqrt{(1 + z_i)(1 + z_j)}$. The spectrograph window function is given by

$$W(k) = e^{-k^2R^2/2}\text{sinc}(k\Delta v/2),$$  \hspace{1cm} (3.12)

where $R$ is the 1σ resolution and $\Delta v$ is the pixel width, both in velocity units.

\(^2\)The pivot point does not matter as long as $v = c \ln \lambda$, and therefore the Ly-$\alpha$ rest-frame wavelength can also be used instead of the median wavelength of the spectrum.
We now move on to defining power spectrum related specifics from data related specifics. In theory, \( w_\alpha(k, z) \) can be any function, but we adopt top-hat \( k \) bands with \( k_n \) as bin edges and linear interpolation for \( z \) bins with \( z_m \) as bin centres: 

\[
w_{(mn)}(k, z) = H(k - k_n)H(k_{n+1} - k)I_m(z),
\]

where \( \alpha \equiv (mn) \), \( H(x) \) is the Heaviside step function and \( I_m(z) \) is the interpolation kernel. We chose the linear interpolation for its smoothness over top hats. However, this mandates distributing pixel pairs into two redshift bins. One can imagine using higher order terms (such as cubic interpolation) to make the function smoother, and hence more accurate, but this will obviously make the calculation more complex. In the end, the linear interpolation is a good compromise between accuracy and complexity.

\[
I_m(z) = \begin{cases} 
\frac{z - z_{m-1}}{z_m - z_{m-1}}, & z_{m-1} < z < z_m \\
\frac{z_{m+1} - z}{z_{m+1} - z_m}, & z_m < z < z_{m+1} \\
0, & \text{otherwise}
\end{cases}
\]  

(3.13)

Note that this is 1 when \( z = z_m \) and 0 when \( z = z_{m\pm 1} \). The derivative matrix for redshift bin \( m \) and wavenumber bin \( n \) is then

\[
Q^{(mn)}_{ij} = I_m(z_{ij}) \int_{k_n}^{k_{n+1}} \frac{dk}{\pi} \cos(kv_{ij})W^2(k).
\]

(3.14)

We compute these matrices for as many redshift bins as necessary for a given spectrum.

Finally, we assume that the noise of every pixel is independent. This results in a diagonal noise matrix with \( N_{ii} = \sigma_i^2 \), where \( \sigma_i \) is the pipeline noise divided by the continuum and the mean normalized flux \( \bar{F}(z) \).

### 3.2.2 Continuum Limit

It is sufficient to take the estimated power as 

\[
P_{\text{est}}(k^n_c, z_m) = P_{\text{fid}}(k^n_c, z_m) + \theta_{(n,m)}
\]

in our analysis, where \( k^n_c \) is the bin centre. However, in order to further improve our intuition, let
us discuss what equation (3.3) constructs in detail. First, we should differentiate between the underlying true power $P_{\text{true}}$ of data and the fiducial power $P_{\text{fid}}$ of the estimator; these two are not necessarily the same. For simplicity, let us ignore redshift dependence, spectrograph resolution and noise, and adopt band powers for $w_n(k) = H(k - k_n)H(k_{n+1} - k)$. In the continuum limit, matrix multiplications can be converted into integrals. Then, equation (3.3) becomes the following at iteration $X$:

$$\theta_n^{(X+1)} = \int_{k_n}^{k_{n+1}} dk \gamma_n^{(X)}(k)[P_{\text{true}}(k) - P_{\text{fid}}(k)],$$

and

$$\gamma_n^{(X)}(k) = \frac{1}{P^2_{(X)}(k)} \left[\int_{k_n}^{k_{n+1}} \frac{dk}{P^2_{(X)}(k)}\right]^{-1},$$

where $P_{(X)}$ is exactly the fiducial power at the first iteration, and approaches to the true underlying power $P_{\text{true}}$ by the last iteration. Thus, this estimator gives us an inverse variance weighted average of the residuals at each iteration. As these residuals gets smaller, the effect of averaging gets smaller as well.

### 3.3 Implementation

QMLE implementation presents challenges in memory, CPU time, numerical stability and confident validation of the results. In this section, we clarify our implementation decisions in order to overcome such challenges.

We will refer Palanque-Delabrouille et al. (2013) as PD13, Walther et al. (2017) as W17 and McDonald et al. (2006) as M06 from now on.

#### 3.3.1 Algorithm

At every iteration, the algorithm for each spectrum is as follows:

1. Compute the covariance matrix using the previous iteration’s $\hat{\theta}$ estimates, then invert
the covariance matrix. Note that the fiducial signal matrix stays fixed.

2. Compute weighted data vector $y = C^{-1} \delta_F$, then $d_\alpha = y^T Q_\alpha y$.

3. Compute $C^{-1} Q_\alpha C^{-1}$, then $b_\alpha$ and $t_\alpha$ using equations (3.5) and (3.6).

4. Compute Fisher matrix using equation (3.7). Note that this needs to consider all redshift bins that the spectrum spans.

We then sum $d_\alpha$, $b_\alpha$, $t_\alpha$ and $F_{\alpha\alpha'}$ of every quasar. Finally, we invert $F$ and find $\hat{\theta}$ estimates using equation (3.3). We check for convergence by comparing these results to the previous iteration using the expressions in section 3.3.3.

One can bootstrap QMLE results easily by saving $F_{\alpha\alpha'}$ and $(d_\alpha - b_\alpha - t_\alpha)$ of each spectrum to a file. Since each spectrum is treated independently, there is no need to recompute these for every realization. One can generate as many bootstrap realizations as needed on spectrum level, and then simply add these saved quantities with repetition to find a bootstrapped estimate. This treatment is exact at the first iteration, and should be a good approximation at convergence.

We find that using a smooth weighted spline in step 1 makes the algorithm numerically stable. This smoothing spline is performed on $(k, P = P_{\text{fid}} + \theta)$, and it can be more reliable if performed on $(\ln k, \ln P)$ while non-positive values are removed.

This algorithm can be implemented using only three matrices for each quasar: One holds the covariance matrix and its inverse, and the other two temporarily hold the derivative and fiducial signal matrices. Holding all matrices in memory for a quasar at a time can improve computation time when memory is the lesser concern, which we find possible in our tests.

This method demands substantial CPU time, and therefore some optimisation mechanisms are noteworthy. First, instead of integrating $S_{\text{fid}}$ and $Q_\alpha$ when needed, we create lookup tables once and interpolate. Moreover, every spectrum can be computed in parallel.
Table 3.1: Top: PD13 BOSS likelihood fitting parameters. Bottom: This work fitting combined PD13 and W17 data with Lorentzian decay added.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>n</th>
<th>α</th>
<th>B</th>
<th>β</th>
<th>$k_1$ [km s$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06</td>
<td>-2.55</td>
<td>-0.10</td>
<td>3.55</td>
<td>-0.28</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>0.066</td>
<td>-2.685</td>
<td>-0.22</td>
<td>3.59</td>
<td>-0.18</td>
<td>0.053</td>
</tr>
</tbody>
</table>

as they are assumed independent. Each CPU should have near-equal workload for efficient parallelization. Matrix operations scale as $O(N_D^3)$, where $N_D$ is the number of pixels in a spectrum. Let us define $N_B$ as the number of total bins to which a spectrum contributes. Then, the Fisher matrix calculation will require $N_B(N_B + 1)/2$ matrix multiplications for that spectrum. Hence, we use $T_{cpu} = N_D^3N_B(N_B + 1)$ as an estimate for computation time, and distribute spectra accordingly.

We use GSL$^3$ for interpolation, integration, matrix inversion; Intel’s MKL library$^4$ for matrix multiplication; and we compile with OPEN-MPI$^5$ for parallelization. Smooth weighted spline is constructed using SCIPY$^6$.

### 3.3.2 Fiducial Power Spectrum

We exploit a baseline estimate of power spectrum in our analysis as discussed in section 3.2. PD13 provides a fitting function with best-fitted parameters. We further modify their fitting function with a Lorentzian decay:

$$
\frac{kP(k, z)}{\pi} = A \frac{(k/k_0)^{3+n+\alpha \ln k/k_0}}{1 + (k/k_1)^2} \left( \frac{1 + z}{1 + z_0} \right)^{B+\beta \ln k/k_0},
$$

(3.17)

where $k_0 = 0.009$ km s$^{-1}$ and $z_0 = 3.0$. However, PD13 measures power spectrum up to 0.02 km s$^{-1}$, so their parameters are not valid on small scales. We combine W17’s power spectrum estimates, and fit the resulting data set. Our modification reduces $\chi^2_\nu$ from 16.5 to

---

$^3$https://www.gnu.org/software/gsl/
$^4$https://software.intel.com/en-us/mkl
$^5$https://www.open-mpi.org
$^6$https://www.scipy.org
5.6. Even though the fit should not be used for scientific purposes, it should be sufficient for a baseline estimate.

Our implementation is also equipped with taking a tabulated fiducial power as input. This feature gives greater freedom in the choice of fiducial. We use this to eliminate any discord between synthetic data and the estimator; and to investigate different choices as fiducial.

### 3.3.3 Convergence

We choose a convergence criterion that summarizes the overall fluctuations between iterations. For this purpose, we calculate the weighted average of the changes between iterations using estimated Gaussian errors. In other words, we define convergence when

\[
\Delta \chi = \sqrt{\frac{1}{N} \sum_{\alpha} \frac{(\Delta \hat{\theta}_\alpha)^2}{F^{-1}_{a\alpha}}} \tag{3.18}
\]

or

\[
\Delta \chi_F = \sqrt{\frac{1}{N} (\Delta \hat{\theta})^T F (\Delta \hat{\theta})} \tag{3.19}
\]

becomes smaller than \(\chi_c = 0.01\), where \(\Delta \hat{\theta}_\alpha \equiv \hat{\theta}_\alpha^{(X+1)} - \hat{\theta}_\alpha^{(X)}\) and \(N\) is the total number of bins. Both expressions return close values and reach convergence at the same iteration in almost all cases.

### 3.3.4 Validation

We use log-normal mocks for validation. These semi-realistic spectra are crucial to examine the accuracy, precision and efficiency of our method.

We generate log-normal mocks by using a modified version of M06. These realizations approximately produce theoretically expected mean flux redshift evolution (Faucher-Giguère et al., 2008; Becker et al., 2013) and power spectra similar to PD13 and W17.
Figure 3.1: The relative error in our measurements from 100 log-normal catalogues where each has 1000 spectra in every redshift bin. Our method reaches sub-percent level precision in the absence of any systematic. Bins go up to the Nyquist frequency.
Figure 3.2: Power spectrum estimates from 800,000 DESI-lite spectra. Error bars from the diagonal of the inverse Fisher matrix. Dotted line in each upper panel is the true power. Noise begins to dominate in grey shaded area, $k > k_N = 0.025$ km/s. The estimates reach percent-level accuracy in lower redshift bins. However, last redshift bins are poorly estimated due to significant decline in statistics. The $\chi^2$ calculated from diagonal (all) elements is 240 (248) for 228 degrees of freedom.
Figure 3.2: Power spectrum estimates from 800,000 DESI-lite spectra continued.
1. Generate a long high-resolution Gaussian random grid with equal spacing in velocity \( v \), zero mean and unit variance.

2. FFT this grid and multiply with \( \sqrt{P(k)/dv} \) to obtain \( \tilde{\delta}_b(k) \), where \( dv \) is the grid spacing in velocity units and the power spectrum is

\[
P(k) = \frac{(k/k_0)^{n-\alpha \ln(k/k_0)}}{1 + (k/k_1)^\gamma},
\]

where \( k_0 = 0.001 \text{ s km}^{-1} \), \( k_1 = 0.04 \text{ s km}^{-1} \), \( n = 0.5 \), \( \alpha = 0.26 \) and \( \gamma = 1.8 \). Inverse FFT and save the variance of this grid \( \sigma^2 \). This is a crude Gaussian base for baryon fluctuations \( \delta_b(v) \) with defined power spectrum at \( z_0 = 3 \).

3. Multiply with a redshift evolution factor \( a(z) \). Such that \( \delta_b(z) = a(z)\delta_b \) and \( \sigma^2(z) = a^2(z)\sigma^2 \).

\[
a^2(z) = 58.6 \left( \frac{1 + z}{1 + z_0} \right)^{-2.82}
\]

4. Apply a squared log-normal transformation to approximate the non-linear and non-Gaussian H\( \text{I} \) column density field.

\[
n(z) = e^{2\delta_b(z) - \sigma^2(z)}
\]

5. Transform this to optical depth \( \tau \) by multiplying with another redshift dependent function.

\[
\tau(z) = 0.55 \left( \frac{1 + z}{1 + z_0} \right)^{5.1} n(z)
\]

6. Finally, the flux is \( F(z) = e^{-\tau(z)} \).

7. Smoothing \( F(z) \) with a Gaussian kernel and re-sampling it onto the observed wavelength grid will result in a spectrograph function in equation (3.12).
Inverting C Matrices
Calculating Q Matrices
Calculating Weighted Q Matrices
Calculating Fisher Matrix

Figure 3.3: Percentage of time spent in different steps for 800,000 DESI-lite spectra. The weighted Q matrices are $C^{-1}Q_\alpha C^{-1}$. Fisher matrix calculation consumes the most time, which also requires calculating $Q_\alpha$ matrices as prerequisite. The time for remaining steps are insignificant.
The mean flux and power spectrum of these mocks can be analytically computed. Using one-point probability of the base Gaussian random field $\delta$, we can write down the following integral for the mean flux:

$$
\bar{F}(z) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left[ -\frac{\delta^2}{2\sigma^2} - x(z)e^{2a(z)\delta} \right] d\delta,
$$

(3.24)

where we have defined

$$
x(z) \equiv 0.55 \left( \frac{1 + z}{1 + z_0} \right)^{5.1} e^{-a^2(z)\sigma^2}.
$$

(3.25)

Although we have found a closed form using saddle point integration (see Appendix B.1), it deviates from the truth for $z \lesssim 3$ up to 4%. It is more accurate to just integrate this expression; and Gauss-Hermite quadrature provides a reliable fast integration.

Power spectrum expression needs two-point probability. We start with flux fluctuations $\delta_F(v, z) = F(v, z)/\bar{F}(z) - 1$, then express the correlation function $\xi_F$ as an integral over two Gaussian random $\delta$s assuming all pixels are at the same redshift. We drop $z$ for clarity.

$$
1 + \xi_F(v_{ij}) = \int \frac{e^{-\delta^T C^{-1} \delta/2} F_i F_j}{2\pi \sqrt{\det C} F^2} d\delta,
$$

(3.26)

where

$$
C = \begin{pmatrix}
\sigma^2 & \xi_G(v_{ij}) \\
\xi_G(v_{ij}) & \sigma^2
\end{pmatrix},
$$

(3.27)

and $\xi_G(v)$ is the correlation function of the base Gaussian field. To convert this expression into Gauss-Hermite quadrature with two variables, we apply Cholesky decomposition.
\[ C = LL^T \] and transform \( \delta = \sqrt{2}L \cdot y \), where

\[
L = \begin{pmatrix}
\sigma & 0 \\
\frac{\xi_G(v)}{\sigma} & \sqrt{1 - \frac{\xi_G(v)^2}{\sigma^2}}
\end{pmatrix}
\]

(3.28)

Here is the final expression for completeness:

\[
1 + \xi_F(v) = \int e^{-\frac{y^2}{2}} \frac{1}{\pi F^2(z)} \exp \left[ -x(z) \left( e^{2a\delta_1} + e^{2a\delta_2} \right) \right] dy
\]

(3.29)

We perform an initial test in the absence of redshift distribution and resolution effects. We start with a long fine grid \( (dv = 0.43 \text{ km s}^{-1}, N = 2^{20}) \), and generate mocks that are discretely distributed in redshift. For validation, we pick four redshifts (2.2, 2.4, 2.6, 2.8), and simulate 100 catalogues where each catalogue has 1000 spectra in every redshift bin. These spectra are exactly centred at these redshifts. We set the spectrograph resolution \( R = 71600 \) to diminish its effect, limit the spectral length to half of the bin size \( \Delta z = 0.1 \), and finally re-sample to pixel size of \( \Delta v = 20.8 \text{ km s}^{-1} \). Figure 3.1 shows our results. Our method reaches sub-percent level accuracy in the absence of any systematic.

### 3.4 DESI-Lite Spectra

To test the feasibility of our method for future DESI spectra, we generate semi-realistic data set using the following simplifying assumptions:

1. We set the observed wavelength grid between 3600–9800 Å. This means the closest forest pixel is at \( z = 1.96 \).

2. We create a logarithmically spaced wavelength grid with \( \Delta v = 30 \text{ km s}^{-1} \). Our grid spacing corresponds to a Nyquist frequency of 0.1 s km\(^{-1}\).

\footnote{DESI will use linear wavelength spacing in its pipeline, but QMLE does not require equal velocity spacing.}
Figure 3.4: The normalized Fisher matrix between redshift bins for $k = 0.0071 \text{ s km}^{-1}$ bin. The off-diagonal terms are approximately 40%. Only neighbouring redshift bins dominate the Fisher matrix, since a pixel pair is split into two bins.

3. We assume a constant resolution power of $R = 3200 (\approx 94 \text{ km s}^{-1} \text{ FWHM in velocity units})$ for all spectra at all wavelengths.

4. We add Gaussian random errors with $\sigma = 0.7$ to $F$, which is approximately S/N$= 2$ Å$^{-1}$.

5. We set the minimum redshift of a Ly-α quasar to 2.1 and the maximum to 4.4, and pick random redshifts from a distribution $n(z)$ (Palanque-Delabrouille et al., 2016; DESI Collaboration et al., 2016). We always limit the forest to [1050 Å, 1180 Å] range in quasar’s rest frame.

6. Finally, we assume DESI will observe 800,000 quasar spectra.

We choose 12 redshift bins between 2.0 and 4.2 with $\Delta z = 0.2$. First 5 $k$ bins are linearly spaced with $\Delta k_{\text{lin}} = 0.001 \text{ s km}^{-1}$, and the following 14 bins are logarithmically spaced with $\Delta k_{\text{log}} = 0.1$. Hence, 0.0005 s km$^{-1} \leq k \leq 0.112$ s km$^{-1}$.
Figure 3.2 shows power spectrum estimates for each redshift bin with error bars from the diagonal elements of the inverse Fisher matrix. The chi-square using the full Fisher matrix is \( \chi^2 / \nu = 248 / 228 \), which implies a valid agreement with the truth. We measure the power spectrum to sub-percent accuracy at lower redshifts, but the accuracy and precision get progressively worse towards high redshifts due to declining quasar numbers. We also expected noise and window function corrections to dominate at high \( k \) as the noise power crosses the signal at \( k_N \approx 0.025 \text{ s km}^{-1} \). This constitutes our foundation as validation of our method.

We would like to stress a subtle point here. First of all, the estimator constructs correct covariance matrices in the first iteration by using the true power as fiducial input. This means we also have the correct Fisher matrix and expect \( \theta \approx 0 \) after the iteration. However, the convergence criteria in equation 3.18 will still yield nearly 1 due to the statistical fluctuations of the power spectrum (note \( \text{Var}[\theta_\alpha] \sim F_{\alpha\alpha}^{-1} \)). Therefore, when the estimator goes into the second iteration, it misidentifies these intrinsic fluctuations as corrections to the covariance matrices and readjusts the Fisher matrix. The smoothing spline ameliorates this digression, though imperfectly.

Fisher matrix calculation consumes the most time as it requires calculating \( Q_\alpha \) matrices and multiplying them \( \mathcal{O}(N^2) \) many times. We identified that at least 46% of the total time goes into the Fisher matrix (see figure 3.3). Moreover, we also found that the Fisher matrix is a band matrix that is prominently tridiagonal for individual \( k \) and \( z \) bins. We normalize the Fisher matrix with respect to its diagonal elements for a clear representation in figure 3.4, which is for \( k = 0.0071 \text{ s km}^{-1} \) bin, but represents a typical redshift dependence\(^10\). This tridiagonal shape is due to keeping the full forest and distributing pixel pairs into two redshift bins. This structure weakly persists between \( k \) bins for a given \( z \) bin as well. Given

\(^8\)We performed 9 independent runs and found \( \chi^2 \) fluctuating around 228 as expected. The value 248 is from our first run, and not a special case.

\(^9\)This \( k_N \) roughly corresponds to \( P_N / \sqrt{N_{\text{qso}}} = P_{1D} \), where the noise power is \( P_N = (\sigma \Delta v / F(z))^2 \) with \( N_{\text{qso}} = 100,000 \) for all redshift bins and \( P_{1D} \) is obtained analytically.

\(^10\)The lowest \( k \) bin is weakly coupled (\( \sim 1\% \)) to an additional redshift bin.
Figure 3.5: The effect of gaps in $P_{1D}$ for two redshift bins $z = 2.2$ and $z = 3.0$. Points labelled "Mask" are corrected measurements by large pixel errors, whereas "Gap" points have unchanged errors. For clarity, we slightly shift $z = 3.0$ points and omit error bars from uncorrected results. The masking error propagates to all scales.

Fisher matrix is the longest step, limiting its calculation to only these terms will speed up the estimation significantly by decreasing the number of matrix multiplications to $O(N)$. We consider this optimisation scheme and how well it performs in section 3.4.4. It is also worth pointing out that the Fisher matrix does not depend on the data, but only on the input power spectrum. Therefore, one could also choose a common Fisher matrix for analysis with many simulations.

We also computed the power spectrum covariance matrix with 5000 bootstrap realizations using the results of the first iteration as discussed in section 3.3.1. We found that the diagonals of the covariance matrix estimated from this bootstrap procedure agreed within percent level with the formal estimates from the Fisher matrix except for the high redshift bin where we do not have enough quasars to compute robust bootstrap errors. This bootstrapping procedure therefore provides a straightforward test of the assumptions underlying the error estimates from the Fisher matrix.
3.4.1 Gaps in Spectra

An advantage of our estimator is that it works in pixel space and can therefore robustly handle missing data in the spectrum. In order to test how masking affects our results, we remove continuous regions from some spectra. We assign 15% probability of having a high-column density absorber (HCD) in a spectrum, matching Noterdaeme et al. (2012). If a spectrum has an HCD, we randomly pick a central wavelength and mask 12.5 Å on each side by setting the flux to the mean value ($\delta_F = 0$). We apply this random masking procedure to 100,000 spectra. We perform a run where masked pixels are removed by assigning large errors, and another run without removing these pixels for comparison.

Figure 3.5 shows our results in two redshift bins. The masking suppresses power at small scales, but propagates to all scales. The run without any correction yields extremely poor results with $\chi^2 \sim 4,000$. When we correct for these pixels by assigning large errors, the power spectrum estimates yield $\chi^2 = 234$, close to the original case. This confirms our method is robust against gaps in spectra.

3.4.2 Continuum Marginalization

In a typical pipeline, the real observable flux $f(\lambda)$ is divided by the quasar continuum $C(\lambda)$ to obtain the normalized flux $F(\lambda)$ for each spectrum. Furthermore, flux fluctuations are obtained by diving this with the mean normalized flux $\bar{F}(\lambda)$. The errors in this process propagate to mostly large scales and are called the continuum errors in Ly-\(\alpha\) nomenclature.

Our quadratic estimator is armed with marginalization capability to suppress these offsets (Slosar et al., 2013, Appendix B). By modifying the covariance matrix to $C' = C + Ntt^T$, where $N$ is large and $t$ is the mode we want to marginalize out, one can show that $C'^{-1}(\delta'_F + \alpha t) \approx C^{-1}\delta'_F$, where the new data vector $\delta'_F$ is orthogonal to $t$. This effectively removes any information from data that is in mode $t$.

We will focus our attention to the continuum fitting Slosar et al. (2013) uses. They
Figure 3.6: Average power spectra of eight runs at $z = 2.2$ and $z = 3.2$ with and without marginalizing continuum errors. The continuum errors highly contaminate the first two $k$ bins at $k = 5 \times 10^{-4}$ s km$^{-1}$ and $k = 1.5 \times 10^{-3}$ s km$^{-1}$. The remaining scales are less affected. By marginalizing two continuum modes, we are able to correctly estimate the power at these scales. Average errors of eight runs at $k = 5 \times 10^{-4}$ s km$^{-1}$ bin increase by 25–45%. The change in the second bin is insignificant.

To fit the flux of each quasar $q$ with $f_q(\lambda_o) = A_q(\lambda_o)C(\lambda_r)\bar{F}(\lambda_o)(1 + \delta_{F,q})$, where $\lambda_o$ is the observed wavelength and $\lambda_r$ is the rest frame wavelength. $C(\lambda_r)$ and $\bar{F}(\lambda_o)$ are determined globally, whereas the function $A_q$ has two free parameters for each quasar:

$$A_q(\lambda) = a_q + b_q \frac{\ln \lambda - \ln \lambda_1}{\ln \lambda_2 - \ln \lambda_1},$$

where $\lambda_{1,2}$ are the beginning and the end of the forest in a given spectrum. These two quasar specific parameters source most of the error assuming the global functions are more robust.

To replicate these continuum errors, we add an error $\eta(\lambda)$ to each quasar’s flux fluctuations $\delta_F(\lambda)$. We limit the form of $\eta(\lambda)$ to the equation above in order to perform a controlled test. We randomly generate the two parameters from a Gaussian distribution with $\sigma = 0.1$. We run eight independent sets of 100,000 spectra.

We add a large constant $N_0$ to all elements of the covariance matrix to marginalize the
amplitude, and \(N_1 t t^T\) to marginalize the slope, where \(t = \ln(\lambda/\lambda_{\text{Ly}a})\).

Figure 3.6 presents the average of eight runs. The continuum errors mostly contaminate large scales. The largest scale \(k = 5 \times 10^{-4} \text{ s km}^{-1}\) bins show big offsets, whereas the following \(k\) bin is less affected. When we marginalize these two continuum terms with \(N_0 = N_1 = 1000\), results go back to the expected values. As a side effect, marginalization increases the error in the first bin by 25–45%.

### 3.4.3 Choice of Fiducial

We already established in the previous sections that using the true power as fiducial yields correct results. In this section, we investigate how different fiducial power influences results by considering two additional cases. First, we run the estimator without any fiducial power. Second, imagining a realistic pipeline, we fit equation 3.17 to these no fiducial results, and employ this best-fitted function as fiducial. In order to run multiple independent realizations, we use eight sub-samples of 100,000 spectra.

We compare \(\chi^2\) for every case using respective Fisher matrices, which can be seen in figure 3.7. True fiducial yields the correct answer in one iteration and is the best estimate as expected. Not using any prior naturally starts away from the truth, but converges to the correct power immediately while being relatively a poor fit. Using the results from no prior to construct a better fiducial decisively outperforms no fiducial case, which we found to hold for all eight runs.

The \(\chi^2\) analysis above weights absolute errors with respective Fisher matrices. The differences in these matrices partially source the \(\chi^2\) trend in figure 3.7. In order to decouple error accuracy from \(P_{1D}\) accuracy, we also calculate all \(\chi^2\)s using the true Fisher matrix. We find this quantity still distinguishes different fiducials as before with values close to the last two iterations of the initial analysis, and stays approximately flat over three iterations. This indicates that QMLE yields good \(P_{1D}\) at the first iteration even for no fiducial case,
Figure 3.7: Comparing average $\chi^2$ of eight runs for three different fiducials using respective Fisher matrices. True fiducial yields the correct answer in one iteration and outperforms other cases as expected. Not using any prior naturally starts away from the truth, but converges to the correct power immediately. Using the results from no prior to construct a better fiducial (re-fitted fiducial) decisively surpasses no fiducial case. Repeating this analysis with true Fisher matrix yields the same conclusion. In that case, $\chi^2$ stays constant over three iterations with values close iteration two and three of this figure. This implies that the first iteration yields good $P_{1D}$ and subsequent iterations correct the error estimates.
but it does not improve at subsequent iterations. Evidently, these iterations mostly correct
the error estimates. This behaviour is reasonable given the weights in equation 3.15 stay
approximately constant across bins, and therefore are unaffected by band power corrections.
The only way to overcome the band power discretization is through integrating a fiducial
power.

To summarize, the key points are as follows: 1) even with no fiducial, the power
spectrum is correctly recovered, 2) a simple iteration quickly gets the correct Fisher matrix,
3) a good fiducial yields better power spectrum and error estimates.

3.4.4 Fisher Matrix Approximation

Each quasar’s Fisher matrix requires $N_B(N_B + 1)/2$ matrix multiplications, where $N_B$ is
the number of total bins to which this spectrum contributes. However, figure 3.4 tells us
that only few elements are dominant in the Fisher matrix. Moreover, we see this structure
in all cases, i.e. different fiducial powers, masking and continuum marginalization.

While the sparsity structure of the Fisher matrix appears to be quite generic, the exact
details (as to the number of non-zero elements) do appear to depend on the total signal-to-
oise of the sample. For the DESI-like data considered in this work, we find that including
a single off-diagonal element in the $k$– and $z$– directions works very well. Specifically, for
every $(k_n, z_m)$ pair we calculate $(k_{n+1}, z_m)$ and $(k_n, z_{m+1})$ elements besides the diagonal.
That reduces the number of matrix multiplications to $3N_B$, which will boost the speed.
Note that this should also be taken into account in load balancing.

We test this optimisation scheme on 800,000 spectra. We find that the resulting $\chi^2$ is
not changed while the average time per iteration decreases by 64% and the time spent in
Fisher matrix calculation per iteration drops by 95%.

Even though we do not have an exact prescription for computing the number of non-zero
elements, the above discussion suggests a procedure by which one starts to fill in the Fisher
Table 3.2: Our model power spectrum and precision forecasts for 5-year DESI survey for fixed $k_f = 2.5 \times 10^{-3}$ s km$^{-1}$. The precision is defined as error divided by signal: $e/P_{1D}$. For reference, metal contamination constitutes 5–15% of the Ly-$\alpha$ power in reality.

<table>
<thead>
<tr>
<th>$z$</th>
<th>$P_{1D}(k_f)$ [km s$^{-1}$]</th>
<th>Precision S/N=1 Å$^{-1}$</th>
<th>Precision S/N=2 Å$^{-1}$</th>
<th>Precision S/N=4 Å$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>10.2</td>
<td>1.31%</td>
<td>0.44%</td>
<td>0.22%</td>
</tr>
<tr>
<td>2.2</td>
<td>13.4</td>
<td>1.08%</td>
<td>0.37%</td>
<td>0.20%</td>
</tr>
<tr>
<td>2.4</td>
<td>17.5</td>
<td>1.12%</td>
<td>0.40%</td>
<td>0.23%</td>
</tr>
<tr>
<td>2.6</td>
<td>22.8</td>
<td>1.24%</td>
<td>0.46%</td>
<td>0.27%</td>
</tr>
<tr>
<td>2.8</td>
<td>29.5</td>
<td>1.46%</td>
<td>0.56%</td>
<td>0.34%</td>
</tr>
<tr>
<td>3.0</td>
<td>37.8</td>
<td>1.87%</td>
<td>0.74%</td>
<td>0.45%</td>
</tr>
<tr>
<td>3.2</td>
<td>48.3</td>
<td>2.50%</td>
<td>0.99%</td>
<td>0.62%</td>
</tr>
<tr>
<td>3.4</td>
<td>61.4</td>
<td>3.34%</td>
<td>1.32%</td>
<td>0.82%</td>
</tr>
<tr>
<td>3.6</td>
<td>77.6</td>
<td>4.82%</td>
<td>1.89%</td>
<td>1.15%</td>
</tr>
<tr>
<td>3.8</td>
<td>97.7</td>
<td>8.67%</td>
<td>3.32%</td>
<td>1.98%</td>
</tr>
<tr>
<td>4.0</td>
<td>123.0</td>
<td>22.00%</td>
<td>8.14%</td>
<td>4.63%</td>
</tr>
<tr>
<td>4.2</td>
<td>153.0</td>
<td>95.30%</td>
<td>33.30%</td>
<td>18.10%</td>
</tr>
</tbody>
</table>

3.4.5 5-Year Forecasts

We would like to make simple predictions for future DESI performance using our simple spectra. We run two additional cases where S/N=1 Å$^{-1}$ and S/N=4 Å$^{-1}$ on 800,000 spectra. In order to speed up our analysis we turn on the Fisher optimisation scheme. We perform only one iteration with true power as fiducial.

We compare $P_{1D}$ estimates for different redshifts at a fixed wavenumber $k_f = 2.5 \times 10^{-3}$ s km$^{-1}$. Our results are summarized in table 3.2. This prediction is under ideal circumstances, hence further complications should be added for accuracy. For example, metal contamination constitutes 5–15% of the Ly-$\alpha$ forest power (Day et al., 2019).

We provide the model power spectrum and the full Fisher matrices in the electronic form. A summary of the Fisher matrices is in table 3.3.

In the noise dominated limit, the covariance matrix is $C \approx N$, and so the Fisher matrix
We would like to start our discussion by highlighting the differences between QMLE and FFT. As we have shown in section 3.2.2, QMLE finds inverse variance weighted averages across bins. Measuring deviations from a baseline power further lessens the averaging, so that QMLE yields near exact $P_{1D}(k)$. On the other hand, FFT estimator computes simple scales as $N^{-2}$. Assuming the noise is uncorrelated and Gaussian with standard deviation $\sigma$, this means that the Fisher matrix is proportional to $F \propto \sigma^{-4} = (S/N)^4$ (or the precision scales as $(S/N)^2$), where we substituted the definition $S/N \equiv 1/\sigma$. We confirmed that this scaling holds true for high $k$ values in the Fisher matrix.

### Table 3.3: Our Fisher matrix forecasts for 5-year DESI survey for different spectral qualities.

The subscript in $F$ refers to the S/N value. The full table and model power spectrum can be found in the electronic submission of this article.

<table>
<thead>
<tr>
<th>$(z_i, k_i)$ (s km$^{-1}$)</th>
<th>$(z_j, k_j)$ (s km$^{-1}$)</th>
<th>$F_1$ [s$^2$ km$^{-1}$]</th>
<th>$F_2$ [s$^2$ km$^{-1}$]</th>
<th>$F_4$ [s$^2$ km$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2.0, 5.00×10$^{-4}$)</td>
<td>(2.0, 5.00×10$^{-4}$)</td>
<td>6.02867×10$^4$</td>
<td>6.51197×10$^4$</td>
<td>3.66215×10$^4$</td>
</tr>
<tr>
<td>(2.0, 5.00×10$^{-4}$)</td>
<td>(2.0, 1.50×10$^{-3}$)</td>
<td>6.63624×10$^0$</td>
<td>6.52238×10$^1$</td>
<td>3.02865×10$^2$</td>
</tr>
<tr>
<td>(2.0, 5.00×10$^{-4}$)</td>
<td>(2.2, 5.00×10$^{-4}$)</td>
<td>2.12509×10$^1$</td>
<td>2.30833×10$^2$</td>
<td>1.33759×10$^3$</td>
</tr>
<tr>
<td>(2.0, 1.50×10$^{-3}$)</td>
<td>(2.0, 1.50×10$^{-3}$)</td>
<td>6.54364×10$^1$</td>
<td>6.00179×10$^2$</td>
<td>2.47973×10$^3$</td>
</tr>
<tr>
<td>(2.0, 1.50×10$^{-3}$)</td>
<td>(2.0, 2.50×10$^{-3}$)</td>
<td>5.99838×10$^0$</td>
<td>5.43360×10$^1$</td>
<td>2.19731×10$^2$</td>
</tr>
<tr>
<td>(2.0, 1.50×10$^{-3}$)</td>
<td>(2.2, 1.50×10$^{-3}$)</td>
<td>1.97400×10$^1$</td>
<td>1.75264×10$^2$</td>
<td>6.90261×10$^2$</td>
</tr>
<tr>
<td>(2.0, 2.50×10$^{-3}$)</td>
<td>(2.0, 2.50×10$^{-3}$)</td>
<td>6.38029×10$^1$</td>
<td>5.69068×10$^2$</td>
<td>2.25086×10$^3$</td>
</tr>
<tr>
<td>(2.0, 2.50×10$^{-3}$)</td>
<td>(2.0, 3.50×10$^{-3}$)</td>
<td>5.83666×10$^0$</td>
<td>5.28085×10$^1$</td>
<td>2.13079×10$^2$</td>
</tr>
<tr>
<td>(2.0, 2.50×10$^{-3}$)</td>
<td>(2.2, 2.50×10$^{-3}$)</td>
<td>1.92028×10$^1$</td>
<td>1.65626×10$^2$</td>
<td>6.24662×10$^2$</td>
</tr>
<tr>
<td>(2.0, 3.50×10$^{-3}$)</td>
<td>(2.0, 3.50×10$^{-3}$)</td>
<td>6.29898×10$^1$</td>
<td>5.69389×10$^2$</td>
<td>2.29742×10$^3$</td>
</tr>
<tr>
<td>(2.0, 3.50×10$^{-3}$)</td>
<td>(2.0, 4.50×10$^{-3}$)</td>
<td>5.74635×10$^0$</td>
<td>5.31461×10$^1$</td>
<td>2.21748×10$^2$</td>
</tr>
<tr>
<td>(2.0, 3.50×10$^{-3}$)</td>
<td>(2.2, 3.50×10$^{-3}$)</td>
<td>1.89681×10$^1$</td>
<td>1.66002×10$^2$</td>
<td>6.39374×10$^2$</td>
</tr>
<tr>
<td>(2.0, 4.50×10$^{-3}$)</td>
<td>(2.0, 4.50×10$^{-3}$)</td>
<td>6.21661×10$^1$</td>
<td>5.77621×10$^2$</td>
<td>2.43005×10$^3$</td>
</tr>
<tr>
<td>(2.0, 4.50×10$^{-3}$)</td>
<td>(2.0, 5.65×10$^{-3}$)</td>
<td>5.85085×10$^0$</td>
<td>5.58077×10$^1$</td>
<td>2.44502×10$^2$</td>
</tr>
<tr>
<td>(2.0, 4.50×10$^{-3}$)</td>
<td>(2.2, 4.50×10$^{-3}$)</td>
<td>1.87450×10$^1$</td>
<td>1.68881×10$^2$</td>
<td>6.78999×10$^2$</td>
</tr>
<tr>
<td>(2.0, 5.65×10$^{-3}$)</td>
<td>(2.0, 5.65×10$^{-3}$)</td>
<td>8.19711×10$^1$</td>
<td>7.90781×10$^2$</td>
<td>3.53432×10$^3$</td>
</tr>
<tr>
<td>(2.0, 5.65×10$^{-3}$)</td>
<td>(2.0, 7.11×10$^{-3}$)</td>
<td>6.07740×10$^0$</td>
<td>6.05452×10$^1$</td>
<td>2.85148×10$^2$</td>
</tr>
</tbody>
</table>

...
averages. For wide enough bins, this does not equal to $P_{1D}(k)$. Therefore, it is important to note that these two methods will not fully agree unless special circumstances are met, because they compute mathematically different quantities.

Systematic errors are the significant uncertainty source in $P_{1D}$ analysis. We have considered three such error sources: 1) high-column density absorbers, 2) metal contamination and 3) continuum fitting. Using our method we showed that HCDs can be masked without further complications. Modelling their contribution can now be solely a theoretical question as they are biased tracers in reality. However, metal contamination is more complicated as they are not easily separable from data like HCDs. Metal contamination estimates will inevitably bring statistical errors; and these errors can be fairly higher than our simple forecasts. Nevertheless, all methods face this challenge. Third, continuum errors can be more complicated than our simple model, but we believe these errors can still be marginalized with careful investigation. Even though marginalization washes out some information, it still enables us to keep the largest scales in $P_{1D}$. Furthermore, DESI will supply plenty of spectra to study and better understand these errors in the future.

Two additional error sources not considered here are uncertainties in the estimates of the resolution and the noise. Although we used a simple Gaussian form of the spectrograph resolution and a diagonal noise matrix, the QMLE formalism allows us trivially extend this to more complicated forms, such as described in Bolton & Schlegel (2010). We defer a complete analysis, including a sensitivity analysis to resolution and noise misestimations to future work.

A disadvantage of our method over a direct FFT method is the additional computational time. However, we demonstrated that these calculations are now practical even for surveys of the scope of DESI, as we were able to perform multiple runs of full DESI-like data in a small cluster of 30 nodes with 24 cores each. We also introduced an optimisation scheme that brings down the CPU time significantly. We think this cost is reasonable given QMLE’s capacity to overcome Ly-α forest specific challenges. Further computational speedups such
as using GPUs are likely possible; we defer these to later work.

### 3.6 Summary

The Ly-α forest has emerged as a unique and competitive tool to investigate the large-scale structure of the universe. This technique can probe cosmological parameters on large scales, while being sensitive to the thermal state of the IGM, neutrino masses and new dark matter models on small scales. This small scale physics enriches the 1D and 3D power spectrum of the Ly-α forest.

In this work, we studied the optimal quadratic estimator (QMLE) for $P_{1D}$ by first formulating its generic and Ly-α specific expressions, offered an analytic continuum limit formula to depict what QMLE calculates and a straightforward way to bootstrap its results. Under the simplest terms, QMLE finds an inverse variance weighted average power for each bin. We then outlined the implementation in detail by providing our step-by-step algorithm. We here underline two of our decisions that mitigates the systematics and numerical instabilities: 1) we fed smoothed estimates to the covariance matrix, and 2) we picked a convergence criteria that weights changes between iterations by the error estimates. Then, we described our synthetic spectra and provided analytic expressions for what they construct. These analytic expressions were crucial to ascertain the performance of QMLE.

We generated DESI-oriented synthetic spectra in order to perform comprehensive tests. These mocks assumed constant resolution ($R = 3200$) and noise at all wavelengths, and limited the quasar redshift range to $z_{\text{qso}} \in [2.1, 4.4]$ with pixel width $c\Delta \ln \lambda = 30 \text{ km s}^{-1}$. A summary of our findings is as follows:

1. Using 800,000 these spectra with S/N=2 Å$^{-1}$, we first showed that the power spectrum could be accurately measured with 1% precision. This number declined towards higher redshifts due to diminishing statistics. This proved the absence of biases in our algorithm.
2. We randomly masked $\Delta \lambda = 25 \, \text{Å}$ region of 15% of spectra, and showed that this masking badly affected the measurement when untreated, and proved QMLE was robust against its effects on 100,000 mocks. As real analyses would mask bad pixels and high density absorbers, this strength is invaluable to $P_{1D}$ analysis.

3. We introduced continuum errors by adding wavelength-dependent error $\eta(\lambda) \sim 10\%$ to flux fluctuations $\delta F$. This error function $\eta(\lambda)$ had two independent parameters for each quasar: amplitude and slope. Using eight independent runs with 100,000 spectra each, we demonstrated that these continuum errors contaminate first two $k$ bins, but QMLE could marginalize out the noisiest modes, and recover these scales.

4. We presented, both analytically and by using 100,000 spectra, how a baseline model improved the accuracy. This fiducial power enables us to integrate over bins and to lessen the discretization of band powers. Our proposed unbiased procedure to find this fiducial power is to first measure the power spectrum without any input, then find the best-fitted analytic function on these results. We showed this feature significantly boosted our measurements.

5. Finally, we found that computation was mostly spent on Fisher matrix calculation. We also found that Fisher matrix had a simple structure, which allowed us to come up with an optimisation scheme that reduced the computation time by 60%.

This work represents an initial step towards analyzing the upcoming Lyman-\(\alpha\) datasets with surveys like DESI. The quadratic estimator formalism allows us to optimally use all the available data (even with varying S/N) and to effectively mitigate systematic errors like those arising from an imperfect continuum estimate. We address several practical issues with implementing a QMLE power spectrum code for the Lyman-\(\alpha\) forest. Future work will use these results and the codes described here to analyze both existing high-resolution data as well as upcoming DESI data.
Chapter 4

Applying the Optimal Estimator to High-Resolution Quasar Spectra

4.1 Introduction

The Ly\(\alpha\) forest technique can map out the matter distribution in vast volumes and at small scales \((r \lesssim 1\,\text{Mpc})\) between \(2 \lesssim z \lesssim 5\). At these redshifts, the structure formation is mildly non-linear, but the physics of the Ly\(\alpha\) forest is further enriched by the thermal state and reionization history of the intergalactic medium (IGM) (Hui & Gnedin, 1997; Gnedin & Hui, 1998). The line-of-sight flux power spectrum \(P_{1D}\) has been at the frontier of constraining new physics including IGM thermal evolution (Boera et al., 2019; Walther et al., 2019), neutrino masses (Croft et al., 1999a; Seljak et al., 2006; Palanque-Delabrouille et al., 2015a,b; Yeche et al., 2017) and the nature of dark matter (Boyarsky et al., 2009; Viel et al., 2013; Baur et al., 2016; Iršič et al., 2017b; Garzilli et al., 2019).

Two categories of \(P_{1D}\) data sets have emerged over the years. The first category contains thousands of low- to medium-resolution spectra obtained by the Extended Baryon Oscillation Spectroscopic Survey (eBOSS) (Dawson et al., 2016) and its predecessors; and the corresponding \(P_{1D}\) estimates (McDonald et al., 2006; Palanque-Delabrouille et al., 2013; Chabanier et al., 2019). The upcoming Dark Energy Spectroscopic Instrument (DESI) (Levi
et al., 2013; DESI Collaboration et al., 2016) aims to obtain approximately one million Ly α quasar spectra. Such large sample sizes can probe large scales to constrain cosmology and neutrino masses, but these data sets are limited by noise and resolution at small scales. The second category contains tens to hundreds of high-resolution, high-S/N spectra obtained by various spectrographs including the High-Resolution Echelle Spectrograph (Vogt et al., 1994, HIRES), the Ultraviolet and Visual Echelle Spectrograph (Dekker et al., 2000, UVES) and X-Shooter spectrograph (Vernet et al., 2011). $P_{1D}$ estimates in this category often have been limited to their respective data sets with 10–100 spectra (Croft et al., 1999b; McDonald et al., 2000; Kim et al., 2004; Viel et al., 2013; Walther et al., 2017; Iršič et al., 2017c; Yeche et al., 2017; Boera et al., 2019; Day et al., 2019). The large-scale modes ($\sim 10$ Mpc) are poorly measured due to large sample variance from the small numbers of quasars in this category, but these spectra can probe extremely small scales ($\sim 100$ kpc) which are crucial to constrain thermal state of the IGM and non-standard dark matter models.

In this work, we measure the small-scale $P_{1D}$ from the largest sample of high-S/N quasars using a combination of three public releases (López et al., 2016; O’Meara et al., 2017; Murphy et al., 2019). This combined sample is seven times larger than Walther et al. (2017), but requires attention when combining the different data sets. We use the optimal quadratic estimator formalism and pipeline described in Karaçaylı et al. (2020) to measure $P_{1D}$ . This approach has a number of advantages: by working in pixel space, it is unbiased by gaps in the spectra, allows weighting by both the pixel-level pipeline noise as well as accounting for sample variance from intrinsic Ly α correlations, and naturally allows combining very different data sets (with different pixel spacings, resolutions etc). As we demonstrate below, the large data set and pipeline result in significant improvements to the precision with which we measure $P_{1D}$ .

This chapter is organized in seven sections. Section 4.2 describes the spectra and the preprocessing steps we take. Section 4.3 details our method with a summary of the pipeline and mean flux measurement. We also review the quadratic estimator and validate the
pipeline on simulated data here. Our results and a discussion on systematics of damped Ly α absorbers, continuum errors and metal contamination can be found in Section 4.4. We reflect on our results and their statistical power in Section 4.5. Finally, we summarize in Section 4.6.

4.2 Data

We use three publicly available data sets in this work:

- Keck Observatory Database of Ionized Absorption toward Quasars (KODIAQ) Data Release 2 (DR2) (Lehner et al., 2014; O’Meara et al., 2015, 2017) is observed by HIRES (Vogt et al., 1994) on the Keck I telescope.

- The Spectral Quasar Absorption Database (SQUAD) DR1 (Murphy et al., 2019) is observed by UVES (Dekker et al., 2000) on the European Southern Observatory’s Very Large Telescope (VLT).

- XQ-100 is observed using the X-Shooter spectrograph (Vernet et al., 2011) under the European Southern Observatory Large Programme “Quasars and their absorption lines: a legacy survey of the high-redshift universe with VLT/XSHOOTER” (López et al., 2016).

4.2.1 KODIAQ

KODIAQ DR2\(^1\) has 300 reduced, continuum-fitted, high-resolution quasar spectra at \(0 < z < 5.3\) with resolving power \(R \gtrsim 36,000\) (Lehner et al., 2014; O’Meara et al., 2015, 2017). The continuum is fitted by hand one echelle order at a time using Legendre polynomials. These high-resolution spectra come in 1.3 or 2.6 km s\(^{-1}\) velocity spacing. We

\(^1\)https://koa.ipac.caltech.edu/workspace/TMP_939bFW_53591/kodiaq53591.html
co-add and resample different observations onto a common 3 km s\(^{-1}\) grid using exposure times as weight (Gaikwad et al., 2020). While this resampling step is not required, it significantly reduces our computational cost while not affecting any of the scales of interest.

### 4.2.2 SQUAD

SQUAD DR1\(^2\) consists of 467 fully reduced, continuum-fitted high-resolution quasar spectra at redshifts 0 < \(z< 5\) with resolving power \(R \gtrsim 40,000\) (Murphy et al., 2019). The continuum fitting consists of an automatic phase and then a manual phase to eliminate the remaining artefacts. These spectra are sampled onto varying 1.3–3.0 km s\(^{-1}\) spaced grids in velocity units. As with KODIAQ, we resample these onto a common 3 km s\(^{-1}\) grid.

There are two further important corrections to the reduced spectra. First, the median seeing is smaller than the slit width for some observations. This results in underestimated nominal resolution values and consequently over-correcting spectrograph resolution. We correct the reported nominal resolution by approximating \(R_{\text{cor}} = R_s/\theta\) only when \(s>\theta\), where \(s\) is the slit width and \(\theta\) is the median seeing, both in arc seconds. This yields 25% correction on average with a maximum of 150%. The net effect on \(P_{\text{1D}}\) is less than 3% even at \(k = 0.1\) s km\(^{-1}\) since the resolution comes into effect at significantly smaller scales.

Second, Murphy et al. (2019) note that their pipeline underestimates the errors in saturated absorption lines. They provide \(\chi^2_{\nu}\) of each pixel about the weighted mean when combining multiple exposures. Following King et al. (2012)'s correction, we apply a median filter of size 5 to \(\chi^2_{\nu}\) and multiply the error with \(\sqrt{\text{median}[\chi^2_{\nu}]}\) if it is greater than 1.

### 4.2.3 XQ-100

XQ-100 contains 100 quasars at redshifts 3.5 < \(z< 4.5\) with resolving power ranging from \(R \sim 4,000–7,000\) and spectra from different arms made available (López et al., 2016). These

\[^2\]https://doi.org/10.5281/zenodo.1345974
spectra are obtained from the ESO database\(^3\). For each arm, the continuum is manually fit by selecting absorption free points. The Ly\(\alpha\) forest falls into VIS and UVB arms. Spectra from VIS arm have a pixel spacing of 11 km s\(^{-1}\), whereas UVB spectra are on 20 km s\(^{-1}\) velocity spaced grids. Given the lower resolution compared with KODIAQ and SQUAD, we do not resample these spectra onto another grid. We also keep the spectra from different arms separate, which allows us to keep resolution correction more accurate. For simplicity, we ignore any correlations between overlapping regions on these two arms.

Because the seeing is smaller than the slit width for most observations, the nominal resolution is similarly underestimated for this set. We correct the resolution for this effect by interpolating the tabulated values\(^4\) only when the seeing is smaller than the slit width (Yeche et al., 2017). We do not extrapolate below the smallest provided slit width value. Even though this yields 30% resolution correction on average and doubles the resolution at maximum, the effect in \(P_{1D}\) is smaller. We see an average of 15% correction for UVB arm and 5% for VIS arm in \(P_{1D}\) at \(k = 0.045\) s km\(^{-1}\), our confidence limit for XQ-100.

### 4.3 Method

#### 4.3.1 Summary of the pipeline

Before processing the spectra, we need to correct the nominal resolution values, identify duplicate quasar observations and mark DLAs. We do not mask metal lines, but subtract a statistical estimate of the metal power using side band regions. First, the nominal resolutions of XQ-100 and SQUAD spectra are corrected for seeings that are less than the slit width as described in their respective sections. Second, assuming a data set does not contain duplicates in itself, we identify quasars within 10" of each other from different sets as

\(^3\)http://telbib.eso.org/detail.php?bibcode=2016A%26A...594A..91L
\(^4\)https://www.eso.org/sci/facilities/paranal/instruments/xshooter/inst.html
duplicates, and pick the spectrum that has the highest S/N per km s\(^{-1}\). Third, we visually identify and remove DLAs with the help of a simple automated DLA finder and given catalogs (see Section 4.4.3).

We then process the spectra as follows:

1. Our analysis region is limited to \(1.7 < z < 4.7\) and \(1050 \text{ Å} < \lambda_{RF} < 1180 \text{ Å}\). We remove pixels that fall outside of these bounds.

2. These spectra are still susceptible to reduction artefacts such as spikes due to continuum normalization near an echelle order, sky subtraction or cosmic rays. We remove these outlier artifacts by eliminating pixels with flux values one median absolute deviation (MAD) outside 0 and 1, and by eliminating pixels with errors 3.5 MAD above the median error. In other words, we keep pixels that satisfy the following criteria:

\[
\begin{align*}
\text{MAD}(F) &< F(\lambda) < 1 + \text{MAD}(F) \quad (4.1) \\
0 &< \sigma(\lambda) < \text{median}(\sigma) + 3.5 \times \text{MAD}(\sigma), \quad (4.2)
\end{align*}
\]

where MAD is computed in the Ly \(\alpha\) region. We prefer median statistics because they are robust against outliers.

3. We co-add multiple KODIAQ observations, then resample these and SQUAD spectra onto a common 3 km s\(^{-1}\) spaced grid. We keep XQ-100 data in its original spacing, and do not co-add UVB and VIS arms.

4. We divide by the best-fit mean flux of the corresponding data set to get \(\delta_F = F/F_1 - 1\).

5. We divide the forest into three equal regions in the rest frame, and split all spectra into these chunks to speed up our calculation and help continuum marginalization. We remove chunks that are shorter than 10\% of the entire forest.
Combining KODIAQ and SQUAD with this process results in 1276 spectral chunks from 464 quasars comprised 186 KODIAQ and 278 SQUAD quasars. We are then left with 74 unique XQ-100 quasars.

### 4.3.2 Mean flux

We measure the mean flux of each data set independently and use the respective best fits as mean flux to calculate $\delta_F = F/F - 1$. As we discuss in Section 4.4, this removes some systematic errors in continuum fitting procedure.

We add Ly $\alpha$ variance to the pipeline error on pixel level using a fiducial power from fitting equation (4.4) to previous measurements and the mean flux from Faucher-Giguère et al. (2008) (FG08). We assign the square of the mean S/N per km s$^{-1}$ of the Ly $\alpha$ region as weight, then simply average pixel values. We do not use inverse variance at the pixel level because the pipeline noise and flux are correlated. Specifically, low flux regions almost always have smaller pipeline error estimates. Error for a given bin is the propagated value of the modified pipeline errors that go into that bin. Note that we do not account for systematics here and therefore underestimate the errors. We then fit Becker et al. (2013) form for $\overline{F}(z) = \exp(-\tau(z))$ to our measurements, where $\tau(z)$ is given by

$$\tau(z) = C + \tau_0 \left( \frac{1 + z}{1 + z_0} \right)^\beta,$$

(4.3)

and $z_0 = 3.5$. Fig. 4.1 shows our mean flux measurements from each data set by using $dz = 0.1$ spaced bins. Table 4.1 has the corresponding best-fitted parameters.

Using the errors we obtain from this analysis, we plot the S/N of each combined data set in Fig. 4.2. Since the systematics are not included, we overestimate the S/N, so this figure is for illustration purposes only. We assume signal mean flux to be from FG08. XQ-100 equally contributes to S/N distribution between $3 \lesssim z \lesssim 4$ even though it has lower resolution.
<table>
<thead>
<tr>
<th>Set</th>
<th>$\tau_0$</th>
<th>$\beta$</th>
<th>$C'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FG08</td>
<td>0.675</td>
<td>3.92</td>
<td>0</td>
</tr>
<tr>
<td>KODIAQ</td>
<td>0.373</td>
<td>5.13</td>
<td>0.18</td>
</tr>
<tr>
<td>SQUAD</td>
<td>0.377</td>
<td>5.54</td>
<td>0.24</td>
</tr>
<tr>
<td>XQ-100</td>
<td>2.000</td>
<td>1.00</td>
<td>-1.43</td>
</tr>
</tbody>
</table>

Table 4.1: Best-fitted parameters to our mean flux measurements of each data set when DLAs are masked.

Figure 4.1: Mean flux measurement of each data set when DLAs are masked. Solid transparent lines show best-fitted results. KODIAQ and SQUAD have deviations at low $z$. FG08 is extrapolated to lower redshifts. We address the discrepancies in Section 4.4.
Figure 4.2: S/N as a function of redshift for each data set, where we define S/N to be $F(z)/\sigma(z)$. The signal mean flux is assumed to be from FG08, $F = F_{\text{FG08}}$. The noise in a redshift bin $\sigma(z)$ is the propagated error when mean flux is estimated, i.e., the error bars in Fig. 4.1. Note we do not account for systematics here, and therefore overestimate the S/N.

4.3.3 Quadratic estimator

Our primary method is the quadratic maximum likelihood estimator (QMLE) (Hamilton, 1997; Tegmark et al., 1997, 1998; Seljak, 1998; McDonald et al., 2006). We refer the reader to Karaçaylı et al. (2020) for details and extensive tests. We also implemented a simple FFT estimator that does not account for noise or resolution as a rough cross-check for our results.

An important feature of our QMLE implementation is estimating deviations from a fiducial power spectrum such that $P(k, z) = P_{\text{fid}}(k, z) + \sum_{m,n} w_{mn}(k, z) \theta_{mn}$, where we adopt top-hat $k$ bands with $k_n$ as bin edges and linear interpolation for $z$ bins with $z_m$ as bin centres (Font-Ribera et al., 2018). Palanque-Delabrouille et al. (2013) provides a fitting function with best-fitted parameters. We further modify their fitting function with a
Lorentzian decay (Karaçaylı et al., 2020):

\[
\frac{kP(k, z)}{\pi} = A \left(\frac{k/k_0}{1 + (k/k_1)^2} \right)^{3+n+\alpha \ln k/k_0} \left(1 + \frac{z}{1 + z_0}\right)^B \ln k/k_0,
\]

(4.4)

where \(k_0 = 0.009\) s km\(^{-1}\) and \(z_0 = 3.0\). Combined with Walther et al. (2017) results, the best-fitted parameters are \(A = 0.066, n = -2.685, \alpha = -0.22, B = 3.59, \beta = -0.16\) and \(k_1 = 0.053\) s km\(^{-1}\). We note that this fit has bad \(\chi^2\) and should not be used for scientific purposes, but it is sufficient for a baseline estimate.

Given a collection of pixels representing normalized flux fluctuations \(\delta F\), the quadratic estimator is formulated as follows:

\[
\hat{\theta}^{(X+1)}_{\alpha} = \sum_{\alpha'} \frac{1}{2} F_{\alpha\alpha'}^{-1}(d_{\alpha'} - b_{\alpha'} - t_{\alpha'}),
\]

(4.5)

where \(X\) is the iteration number and

\[
d_{\alpha} = \delta F^T C^{-1} Q_{\alpha} C^{-1} \delta F, \tag{4.6}
\]

\[
b_{\alpha} = \text{Tr}(C^{-1} Q_{\alpha} C^{-1} N), \tag{4.7}
\]

\[
t_{\alpha} = \text{Tr}(C^{-1} Q_{\alpha} C^{-1} S_{\text{fid}}), \tag{4.8}
\]

where the covariance matrix \(C \equiv \langle \delta F \delta F^T \rangle\) is the sum of signal and noise as usual, \(C = S_{\text{fid}} + \sum_{\alpha} Q_{\alpha} \theta_{\alpha} + N, Q_{\alpha} = \partial C/\partial \theta_{\alpha}\) and the estimated Fisher matrix is

\[
F_{\alpha\alpha'} = \frac{1}{2} \text{Tr}(C^{-1} Q_{\alpha} C^{-1} Q_{\alpha'}). \tag{4.9}
\]

The covariance matrices in the right hand side of equation (4.5) are computed using parameters from the previous iteration \(\hat{\theta}^{(X)}_{\alpha}\), assuming different quasar spectra are uncorrelated, the Fisher matrix \(F_{\alpha\alpha'}\) and the expression in parentheses in equation (4.5) can be computed for each quasar, then accumulated, i.e. \(F = \sum_q F_q\) etc.
In order to adapt this to Ly$\alpha$ forest analysis, we first convert a pixel’s wavelength to velocity using logarithmic spacing as has been the cosmology independent convention,

$$v_i = c \ln \left( \frac{\lambda_i}{\lambda_{Ly\alpha}} \right)$$

$$z_i = e^{v_i/c} - 1,$$  \hspace{1cm} (4.10)

where $\lambda_{Ly\alpha} = 1216 \text{ Å}$.

Second, since the resolution matrix is not provided, we make the approximation that the resolution does not change with wavelength and is Gaussian for the rest of the chapter. Then, the signal is the power spectrum multiplied with the spectrograph window function $W(k)$ in Fourier space.

$$S_{ij}^{\text{fid}} = \int_{\infty}^{0} \frac{dk}{\pi} \cos(kv_{ij})W^2(k)P_{\text{fid}}(k, z_{ij}),$$  \hspace{1cm} (4.12)

where $v_{ij} \equiv v_i - v_j$ and $1 + z_{ij} \equiv \sqrt{(1 + z_i)(1 + z_j)}$. The spectrograph window function is given by

$$W(k) = e^{-k^2R^2/2}\text{sinc}(k\Delta v/2),$$  \hspace{1cm} (4.13)

where $R$ is the 1σ resolution and $\Delta v$ is the pixel width, both in velocity units. The derivative matrix for redshift bin $m$ and wavenumber bin $n$ is

$$Q_{ij}^{(mn)} = I_m(z_{ij}) \int_{k_n}^{k_{n+1}} \frac{dk}{\pi} \cos(kv_{ij})W^2(k),$$  \hspace{1cm} (4.14)

where $I_m(z)$ is the interpolation kernel which is 1 when $z = z_m$ and 0 when $z = z_{m\pm1}$. We compute these matrices for as many redshift bins as necessary for a given spectrum.

Finally, we assume that the noise of every pixel is independent. This results in a diagonal noise matrix with $N_{ii} = \sigma^2_i$, where $\sigma_i$ is the continuum normalized pipeline noise divided by the mean normalized flux $\overline{F}(z)$. 

84
Figure 4.3: Power spectrum estimates from one mock set. Error bars are replaced with bootstrap estimates.
4.3.4 Validation

Given the resolution and S/N diversity of KODIAQ, SQUAD and XQ-100 data, it is crucial to verify the power spectrum estimates are unbiased and the errors are correctly estimated in an ideal statistical limit. To validate our method in these both aspects, we generate 100 independent log-normal mock data sets with exact redshift distribution, resolution and noise properties of the data using the procedure described in Karaçaylı et al. (2020). These synthetic spectra approximately produce expected mean flux redshift evolution (Faucher-Giguère et al., 2008) and power spectrum similar to Palanque-Delabrouille et al. (2013) and Walther et al. (2017). Even though these mocks cannot capture all the richness of data, they form the baseline with which we validate QMLE. We use bootstrap error estimates to partially capture effects not present in the mocks. We also discuss some astrophysical and instrumental effects in Section 4.4.6.

We use the true values for fiducial power spectrum and mean flux to estimate $P_{1D}$ with one iteration only. Fig. 4.3 shows a sample result from one mock set. From the average of these 100 estimates, we find that our results are unbiased in the range of interest $k < 0.1 \text{ s km}^{-1}$, but they diverge from the truth for $k \gtrsim 0.2 \text{ s km}^{-1}$ as Fig. 4.4 shows. We note that these scales are noise-dominated as seen in Fig. 4.5, and further complicated by resolution effects and narrow metal lines in real data. Therefore, they are significantly hard to measure and have been absent in previous studies.

We then investigate the accuracy of estimated errors by a $\chi^2$ analysis, and set the average of 100 results as the truth. First, we calculate the $\chi^2$ of each result using the Fisher matrix from QMLE$^5$, which yields larger values than expected. We identify the cause to be the violation of QMLE Gaussianity assumption at small scales by generating Gaussian mocks, which yields the correct $\chi^2$ distribution. Therefore, we get another estimate of the covariance matrix by generating 25 000 bootstrap realizations from one set. We find that

$^5$We note that for all sets QMLE gives the same Fisher matrix, and so the same covariance matrix.
Figure 4.4: Average of 100 mock results divided by the true power. We find that our results are unbiased for $k < 0.1 \text{s km}^{-1}$ from this ratio, but they diverge from the truth for $k \gtrsim 0.2 \text{s km}^{-1}$.
Figure 4.5: Noise power spectrum from one mock set divided by the true power. Noise starts to dominate at $k \approx 0.2 - 0.3$ s km$^{-1}$. 

Figure 4.6: To validate our method, we performed $\chi^2$ analysis on 100 mock sets using the Fisher matrices from QMLE and 25 000 bootstrap realizations from one set. We regularized the bootstrap Fisher matrix by exploiting its sparsity pattern and by flooring the eigenvalues to their Gaussian limits, which is noted as "Reg. Bootstrap". This produces the expected $\chi^2$ distribution.
the diagonals of the bootstrap covariance matrix converges rapidly, but the off-diagonal terms remain noisy (and probably correlated) even with 250,000 realizations. We believe our sample size of approximately 1500 chunks is not enough for the degrees of freedom in the covariance matrix, which is a 315x315 matrix, or for its condition number $O(10^{11})$. To achieve a stable covariance matrix, we apply a two-step regularization scheme.

After generating bootstrap realizations with spectral chunks, our final algorithm is based only on data and as follows:

1. To prevent off-diagonal noise from leaking, we directly estimate the Fisher matrix element-wise, where $|r_{ij}^{\text{QMLE}}| > 0.01$ and $r_{ij} \equiv F_{ij}/\sqrt{F_{ii}F_{jj}}$ using the algorithm in Padmanabhan et al. (2016). This algorithm further refines the element-wise estimate to the "closest" positive-definite matrix.

2. We then find the eigenvalues $\lambda_i$ and eigenvectors $e_i$ of this Fisher matrix. We calculate the precision of these eigenvectors under Gaussianity: $\lambda_i^{\text{QMLE}} = e_i^T F^{\text{QMLE}} e_i$. This is the theoretical maximum (minimum for the covariance), so we replace $\lambda_i \rightarrow \min(\lambda_i, \lambda_i^{\text{QMLE}})$ and rebuild the Fisher matrix (McDonald et al., 2006).

We note that using the bootstrap eigenvectors as basis in step 2 captures the non-Gaussian mode couplings. However, bootstraps are also missing fluctuations in certain modes due to low statistics; for example, there are only 5 chunks in the last redshift bin. This eigenvalue regularization scheme takes these modes to their Gaussian limit, and is a slight modification of McDonald et al. (2006).

Fig. 4.6 shows that the regularized bootstrap Fisher matrix produces the expected $\chi^2$ distribution. These tests give us the confidence that our power spectrum estimates are unbiased, and the covariance can be estimated using bootstrap method at the scales of interest.

We tested our regularization scheme on the bootstrap covariance matrix and observed a marginally worse performance. The element-wise estimate is not positive-definite, but
eigenvalue regularization fixes that problem. However, the Fisher matrix has a simpler structure than the covariance matrix. Furthermore, the $\chi^2$ calculation is a linear operation with the Fisher matrix elements which further justifies the element-wise approximation, whereas inverting the covariance matrix is highly non-linear. Therefore, we opted for a direct estimation of the Fisher matrix in this work.

We also test our mean flux measurement process with one mock set. We find that we obtain the correct mean flux, but the propagated errors are notably underestimated even with added Ly $\alpha$ variance, which can be suspected from Fig. 4.1 as well. We remind the reader that we are not interested in a precise mean flux measurement, but only in using it to remove some continuum fitting systematics from spectra. However, we perform another test to quantify the underestimation of these errors. We resample the spectra onto a coarse grid (300 km s$^{-1}$) using inverse variance (note signal and errors are not correlated for the mocks). We then estimate the standard deviation using these coarse pixels in a given redshift bin, and find that they are four times the propagated errors. We also note that the

Figure 4.7: The normalized bootstrap covariance matrix of data for $k = 0.013$ s km$^{-1}$ bin. Neighbouring redshift bins are anti-correlated because of assigning pixel pairs into two redshift bins.
correlations between pixels actually matter for the resolution and S/N of our data when resampling (Slosar et al., 2013), but it is still unclear if they could yield correct error bars when propagated. For the measurement from data however, the systematics are most likely a significant source of error.

4.4 Results

We measure the power spectrum in 15 redshift bins from \( z = [1.8, 4.6] \) with \( dz = 0.2 \) spacing for 21 \( k \) bins of which the first 4 are linearly spaced with \( \Delta k_{\text{lin}} = 0.0022 \text{ s km}^{-1} \) and the rest are logarithmically spaced with \( \Delta k_{\text{log}} = 0.1 \).

To remind the reader our pipeline, we mask DLAs by using the accompanying catalogs and visually identifying remaining absorbers with damping wings. We do not mask metal lines, but provide a statistical estimate of the metal contamination using side band regions. Furthermore, we marginalize out the constant and the slope (\( \ln \lambda \)) terms of the continuum errors. Because the continuum is fitted piece-wise and corrected manually, these are not the exact expressions for the continuum errors, but the marginalization should still remove some or most of the contamination in the largest scales. These points and their contribution to systematic error budget are further discussed below.

We perform an initial run on the combined set with the fiducial power spectrum as described in Section 4.3. We fit the results for \( k < 0.1 \text{ s km}^{-1} \) to get a new estimate for the fiducial power, and then perform one iteration. The new parameters are \( A = 0.084, n = -2.655, \alpha = -0.155, B = 3.64, \beta = 0.32 \) and \( k_1 = 0.048 \text{ s km}^{-1} \). We showed this process yields better power spectrum estimates in Karaçaylı et al. (2020).

The estimated errors from QMLE are under Gaussian assumption and significantly underestimated at small scales and high redshifts. We estimate the error bars by generating bootstrap realizations and using our regularization scheme as discussed in Section 4.3.4. Since our method divides pixel pairs into two redshift bins, the power spectrum estimates
Figure 4.8: The raw power spectrum measurements of KODIAQ, SQUAD and XQ-100 divided by the raw power from KODIAQ + SQUAD ($P_{KS}$) when DLAs are masked. Error bars are from bootstrap. We highlight the large-scale modes $k \lesssim 4 \times 10^{-3}$ s km$^{-1}$ that are susceptible to continuum errors in light blue, and small-scale modes $k \gtrsim 0.1$ s km$^{-1}$ that are affected by noise, metal contamination and resolution effects in light orange. Light pink region is excluded from our XQ-100 analysis.
Figure 4.9: Our raw and side band subtracted combined power spectrum measurements when DLAs are masked. The error bars are from bootstraps; and systematic error budget is added in quadrature. We highlight the large-scale modes $k \lesssim 4 \times 10^{-3}$ s km$^{-1}$ that are susceptible to continuum errors in light blue, and small-scale modes $k \gtrsim 0.1$ s km$^{-1}$ that are affected by noise, metal contamination and resolution effects in light orange. We remind the reader that the last two $z$ bins have low statistics as well as further continuum fitting complications.
are correlated between these redshift bins. Fig. 4.7 shows an example normalized covariance matrix from bootstrap estimation for $k = 1.3 \times 10^{-2}$ s km$^{-1}$ bin.

### 4.4.1 Consistency between data sets

Our aim in this work is to utilize the combined data for the best S/N, but first we compare individual power spectrum estimates to check our treatment of the data and consistency between data sets. Fig. 4.8 lays out the ratios of individual power spectra to our raw measurement from KODIAQ + SQUAD (KS).

The first notable difference is between KODIAQ and SQUAD at $z = 1.8$, but they both still fall within the large errors of Walther et al. (2017). Otherwise, these two sets agree with each other. To quantify the agreement between them, we calculate $\chi^2$ from the power difference:

$$\chi^2 = d^T (C_1 + C_2)^{-1} d,$$

where $d = P_1 - P_2$. This yields $\chi^2 = 218$ for 180 degrees of freedom (dof) while removing $z = 1.8$ and $k > 0.1$ s km$^{-1}$ bins. This value is $2\sigma$ away from the mean; or in other words, probability of getting $\chi^2 > 218$ is 3%. This is a reasonable probability when comparing two data sets, so we conclude they are consistent and remove $z = 1.8$ from our conservative range$^6$.

We find that our XQ-100 results agree with Iršič et al. (2017c) in their provided range. The largest inconsistency here most visibly comes from $k > 0.06$ s km$^{-1}$ values. We remind the reader that we corrected the resolution for seeing to the best of our knowledge, but it can still carry some errors especially at this range given the Nyquist frequency for XQ-100 is $k \sim 0.1$ s km$^{-1}$. So, accurately measuring these modes is an ambitious goal for

---

$^6$We have a number of hypotheses for this discrepancy. First, the Ly $\alpha$ forest at low redshifts will be in the UV where the response of most spectrographs rapidly falls. Also, metal contamination would have the largest effect at these redshifts and is hard to measure for these samples. However, we were not able to isolate a particular cause, which is why we exclude this redshift bin from our final results.
XQ-100 alone. Given our limited understanding of the resolution, we assign a conservative 
\( k < 0.045 \text{s km}^{-1} \) range for our XQ-100 measurement, where \( \chi^2 = 83 \) for 77 dof when 
systematic errors are included. Unfortunately, this makes combining XQ-100 at the chunk 
level complicated as QMLE in its current form cannot account for systematic errors. We 
defer how these systematic errors can be incorporated into QMLE to a future work. Instead, 
we provide a separate measurement and a covariance matrix for XQ-100 in this work.

### 4.4.2 Final combined results

Let us start this section with a review of the weighted average and its error propagation. 
Assume for a data set \( m \), we have measurements \( x_m \) with respective errors \( C_m \). Using 
normalized weights \( W_m \), the weighted average \( x_W \) and resulting error \( C_W \) are given by

\[
x_W = \sum_m W_m x_m, \quad (4.16)
\]
\[
C_W = \sum_m W_m C_m W^T_m + \sum_{m \neq n} W_m C_{mn} W^T_n, \quad (4.17)
\]

where \( C_{mn} \) is the cross-covariance between measurements \( m \) and \( n \). These become 
the usual inverse variance weighted average expressions with \( W_m \propto (C_m)^{-1} \) if the 
measurements are uncorrelated. In our case however, some systematic errors (DLAs, 
continuum and metals) between KS and XQ-100 are data-independent and correlated, 
whereas resolution systematics for example are not. Let us define \( \Lambda_{\text{cor}} \) to be these data-
independent, correlated systematics between KS and XQ-100, and define \( \Lambda_{m\text{uncor}} \) to be 
the data-dependent, uncorrelated systematics for measurement \( m \). Then, we substitute 
\( C_{mn} = \Lambda_{\text{cor}} \) and \( C_m = C_{m\text{stat}} + \Lambda_{m\text{uncor}} + \Lambda_{\text{cor}} \), and correlated systematic comes out since 
\( \sum_m W_m \equiv 1 \).

\[
C_W = \sum_m W_m (C_{m\text{stat}} + \Lambda_{m\text{uncor}}) W^T_m + \Lambda_{\text{cor}} \quad (4.18)
\]
In our case, \( \Lambda_{\text{cor}} = \Lambda^{\text{DLA}} + \Lambda^{\text{Cont}} + \Lambda^{\text{Metal}} \) and \( \Lambda_{m_{\text{uncor}}} \) is the resolution systematics. In short, uncorrelated systematics will become smaller when different measurements are combined, while correlated systematics stay the same.

We now use the inverse variance weights with systematic errors, which also prevents underestimated statistical errors from dominating the average:

\[
W_m \propto \left( C_{m_{\text{stat}}} + \Lambda_{m_{\text{uncor}}} + \Lambda_{m_{\text{cor}}} \right)^{-1}
\]

We again generate 25,000 bootstrap realizations to calculate the statistical Fisher matrix. Our systematic error budget is the same for KS and XQ-100 measurements except for the resolution (see Section 4.4.6) and is added in quadrature to the diagonal. We also add large numbers to the diagonal of XQ-100 covariance at \( k > 0.045 \) s km\(^{-1}\) to remove these modes from the average. For both measurements, we also subtract the metal power and add its covariance first, so that the weights have all the statistical and systematic errors in place.

Our results from this weighted average are shown in Fig. 4.9, which is in good agreement with previous measurements even at the largest scales (Iršič et al., 2017c; Walther et al., 2017; Chabanier et al., 2019). We caution the reader that the number of quasars and different continuum treatments limit the accuracy of our measurement for \( k \lesssim 4 \times 10^{-3} \) s km\(^{-1}\); and refer the reader to eBOSS (Chabanier et al., 2019) for a better measurement at these scales. Nevertheless, we are encouraged by this agreement at the large scales.

### 4.4.3 Damped Lyman-alpha absorbers

The net result of damped Lyman-alpha absorbers (DLA) on the power spectrum comprises three effects. The primary effect is the large amount of power added to the largest scales due to their damping wings. The central region of complete absorption has two competing effects: suppression of power and amplification because of lower mean flux measurement (McDonald et al., 2005; Rogers et al., 2018). These high-column density absorbers are
Figure 4.10: The ratio of the measured power spectrum from the combined data set between DLAs kept $P_{\text{Kept}}$ and DLAs masked $P_{\text{Masked}}$. The damping wings add power to large scales, and the lower mean flux overcomes the full absorption suppression from medium to small scales by changing it $\approx 5\%$. 

difficult to model and simulate, so they are often removed from spectra. We describe our DLA identification and removal in this section.

The SQUAD data set comes with a DLA catalog; and the ones in XQ-100 are identified in Sánchez-Ramírez et al. (2016). We mask a reported DLA at $z_{\text{abs}}$ between $[\lambda_C - W/2, \lambda_C + W/2]$, where $\lambda_C = (1 + z_{\text{abs}}) \lambda_{\text{Ly}\alpha}$ and $W$ is the equivalent width given by the following equation (Mo et al., 2010, Sec. 16.4.4):

$$W = 7.3(1 + z_{\text{abs}}) \sqrt{\frac{N_{\text{HI}}}{10^{20} \text{ cm}^{-2}}} \, \text{Å}$$  \hspace{1cm} (4.20)

We also apply a simple automated DLA finder which first finds regions that are consecutively $F(\lambda) < F(\lambda) + \sigma(\lambda)$. If a region is longer than the equivalent width for an absorber with $N_{\text{HI}} = 10^{19} \, \text{cm}^{-2}$ in equation (4.20) at the central redshift, we mark it as a DLA candidate. We then visually inspect all these candidates for damping wings to remove
Figure 4.11: The ratio of the measured power spectrum without continuum marginalization $P_{\text{No Mar}}$ to with marginalization $P_{\text{Mar}}$. Continuum errors significantly contaminate the largest scales; the rest fluctuates around less than one percent.

Our results in Fig. 4.9 are already without DLAs. We performed a run where we kept in the DLAs as well. Fig. 4.10 shows the effect described in the beginning of this section. The presence of DLAs significantly affects the largest scales $k \lesssim 3 \times 10^{-1} \text{ s km}^{-1}$, while boosting the intermediate scales by few percent because of lower mean flux estimates.

### 4.4.4 Continuum

The observed flux $f(\lambda)$ is divided by the quasar continuum $C(\lambda)$ and the mean normalized flux $\bar{F}(\lambda)$ to obtain the flux fluctuations $\delta_F(\lambda)$. Since $C(\lambda)$ is smooth, errors in this process propagate to mostly large scales and are called the continuum errors.

We start our discussion by considering a systematic scaling of the continuum in a given data set. In other words, we assume the fitted continuum of a quasar $q$ is the true continuum times a constant, $C_q^*(\lambda) = a C_q(\lambda)$, for all quasars in the set. This inevitably scales the
normalized flux by the same factor, $F_q^* = f_q / C_q^* = F_q / a$, and hence the measured mean flux $\bar{F}^* = \bar{F}/a$ since $a$ does not depend on $q$. Therefore, using the measured mean flux $\bar{F}^*$ of a given data set removes this systematic bias from the fluctuations $\delta_{F,q}^* = F_q^* / \bar{F}^* - 1 = \delta_{F,q}$.

In fact, we see this type of measured mean flux scaling across data sets, where KODIAQ has 6% and XQ-100 has 15% excess mean flux on average compared to SQUAD. We are further encouraged by the agreement of $P_{1D}$ without any scaling between data sets.

Finally, note that this multiplicative scaling would cancel even if it depended on observed wavelength $a = a(\lambda)$ as long as $a$ does not change from quasar to quasar.

Now, let us introduce a quasar dependent error $\eta_q(\lambda)$ that remains after this systematic scaling, i.e. $C_q^* = aC_q[1 + \eta_q(\lambda)]$. Since the continuum itself is smooth, $\eta_q(\lambda)$ will also be smooth and slowly changing given a well-behaving continuum fitting procedure. The effect of this quasar dependent error will propagate to the mean flux measurement by some average.

$$\bar{F}^* = \frac{\bar{F}}{a} \left\langle \frac{1}{1 + \eta_q} \right\rangle_q \approx \frac{\bar{F}}{a} (1 - \bar{\eta}),$$  \hspace{1cm} (4.21)

where we have assumed $\eta_q$ is small and not correlated with $F_q$. Then,

$$\delta_{F,q}^* = \frac{F_q^*}{F^*} - 1 = \frac{F_q / \bar{F}}{(1 + \eta_q)(1 - \bar{\eta})} - 1$$ \hspace{1cm} (4.22)

$$= (1 + \delta_{F,q})(1 - \Delta\eta_q) - 1.$$ \hspace{1cm} (4.23)

Finally,

$$\delta_{F,q}^*(\lambda) = \delta_{F,q}(\lambda)[1 - \Delta\eta_q(\lambda)] - \Delta\eta_q(\lambda),$$ \hspace{1cm} (4.24)

where we have defined $\Delta\eta_q(\lambda) \equiv \eta_q(\lambda) - \bar{\eta}(\lambda)$. Unfortunately, this error does not fully cancel, but it is possible to partially marginalize out the additive term $\Delta\eta_q(\lambda)$ by approximating its form. This error is described by a constant shift and a slope in eBOSS pipeline,
\[ \Delta \eta_q(\lambda) = a_q + b_q \ln \lambda. \] We marginalize out these modes in this work as well, even though they are not the exact expressions for the continuum errors because the continuum is fitted piece-wise and then corrected manually. Dividing the forest into three chunks also helps by limiting the wavelength range of \( \Delta \eta_q(\lambda) \) that needs to be described. Note that since \( \Delta \eta_q(\lambda) \) is smooth, this error affects mostly large scales even when uncorrected. We ignore the multiplicative term within our approximation. Its major complication would stem from the correlations between \( \eta_q(\lambda) - F_q(\lambda) \) and \( \eta_q(\lambda) - \eta_q(\lambda') \), which can ultimately be treated by smaller errors or by more uniform continuum fitting procedures.

We performed another run where we turned off the continuum marginalization to show its effect. Fig. 4.11 shows the largest scales \( k \lesssim 3 \times 10^{-3} \text{ s km}^{-1} \) are again significantly affected by the continuum. However, the intermediate scales deviate by less than one percent.

### 4.4.5 Side band power

The red side of the Ly \( \alpha \) line in the spectrum is free from H \( \text{i} \) absorption, so this region is used to estimate the power from metals and other systematics (McDonald et al., 2006; Palanque-Delabrouille et al., 2013; Chabanier et al., 2019). We define the first side band (SB 1) region to be between 1268 – 1380 Å, below the Si \( \text{IV} \) line, in quasar’s rest frame, and the second side band region to be between 1409 – 1524 Å, below the C \( \text{IV} \) line. We use \( \overline{F}(z) = 1 \), and estimate the power on the same \( k \) bins. We note that this only removes power due to metals with \( \lambda_{RF} \gtrsim 1400 \text{ Å} \), and hence some metal contamination still remains and produces oscillatory features such as Si \( \text{III} \)-Ly \( \alpha \) cross correlation (McDonald et al., 2006; Palanque-Delabrouille et al., 2013).

We limit the side band power estimate to KODIAQ and SQUAD data sets to not further complicate analysis. The final metal power is estimated from the Si \( \text{IV} \) region (SB 1), and uses 1391 chunks with 221 KODIAQ and 271 SQUAD quasars.
Figure 4.12: SB 1 power spectrum estimated with FFT and QMLE.
Figure 4.13: SB 2 power spectrum estimated with FFT and QMLE.
Figure 4.14: Comparison of both side band power spectra estimated with QMLE. The two side band powers mostly agree with each other with some offset at $z = 2.0, 3.2$ and $3.6$ bins, while $z = 2.2$ has a larger gap.
Figure 4.15: The power spectra measured from the two side bands at $z = 3.0$. The C IV doublet causes the oscillation peaks at $k = 1.3 \times 10^{-2}$ s km$^{-1}$ and $2.5 \times 10^{-2}$ s km$^{-1}$. This doublet and others are apparent in the correlation function (see Fig. 4.16).

The initial runs for both side band regions use 10% of the fiducial power in Section 4.3. We then switch to the best-fitted parameters for $k < 0.1$ s km$^{-1}$ as the new fiducial and estimate side band powers with one iteration. These new parameters are $A = 0.0027, n = -2.92, \alpha = -0.174, B = 0.236, \beta = -0.01$ where the Lorentzian term can be ignored as $k_1 \sim 10^{10}$ s km$^{-1}$. We again estimate the covariance matrix using 25 000 bootstrap realizations and use these instead of QMLE errors. Fig. 4.12 and 4.13 compares QMLE estimates to FFT as a cross check. We do not subtract noise or correct for resolution in our FFT estimates. Fig. 4.14 compares both side band powers from QMLE at all redshift bins. We find that the two side band powers mostly agree with each other with some offset at $z = 2.0, 3.2$ and $3.6$ bins, while $z = 2.2$ has a larger gap. Fig. 4.15 shows and compares the side band powers to the Ly $\alpha$ power at $z = 3.0$.

Noteworthy features in both side band powers are the two visible peaks due to oscillations in all redshift bins at $k = 1.3 \times 10^{-2}$ s km$^{-1}$ and $2.5 \times 10^{-2}$ s km$^{-1}$. While bootstrap method shows that the QMLE errors are underestimated for $z < 3$ and overestimated for
Figure 4.16: The residual correlation functions for SB 1 (green squares) and SB 2 (brown triangles). Curves with solid lines are our fitting model for known metal doublets. Our simple fitting function captures the important features. Note again that Si IV line is outside of SB 2, so there is no peak at $v \approx 2000$ km s$^{-1}$.

For $z > 3$, these peaks remain visible in QMLE error to bootstrap error ratio. We find these oscillations are due to C IV doublet, which is manifested as a peak in the 1D correlation function at $v \approx 500$ km s$^{-1}$ separation. Moreover, we identified more doublet features in the 1D correlation function with a simple model. First, we estimate the correlation function by inverse FFTing the power spectrum before binning into $k$ bins. This yields the correlation function on the full-resolution grid, which we then bin into $v$ bins. Doublets of the same absorber manifest as peaks in this correlation function. To increase S/N, we average all redshift bins.

We found that the correlation function has a smooth component, which we model as follows:

$$\xi_{\text{smooth}}(v) = C + \frac{x_0}{1 + (v/v_0)^\gamma}, \quad (4.25)$$

where $C, x_0, v_0$ and $\gamma$ are fitting parameters. We then model the peaks as Gaussian functions:
<table>
<thead>
<tr>
<th>Doublet</th>
<th>$\lambda_1 - \lambda_2$ [Å]</th>
<th>$v$ [km s$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>C IV</td>
<td>1548.20 - 1550.77</td>
<td>497.2</td>
</tr>
<tr>
<td>Mg II</td>
<td>2796.35 - 2803.53</td>
<td>768.7</td>
</tr>
<tr>
<td>Fe II</td>
<td>2374.46 - 2382.76</td>
<td>1046.6</td>
</tr>
<tr>
<td>Fe II'</td>
<td>2586.65 - 2600.17</td>
<td>1563.2</td>
</tr>
<tr>
<td>Al III</td>
<td>1854.72 - 1862.79</td>
<td>1302.2</td>
</tr>
<tr>
<td>Si IV</td>
<td>1393.76 - 1402.77</td>
<td>1932.8</td>
</tr>
</tbody>
</table>

Table 4.2: Doublets used in SB correlation function analysis.

\[
\xi_p(v) = A_p e^{-(v-\mu_p)^2/2\sigma_p^2},
\]  

(4.26)

where $A_p$ and $\sigma_p$ are free parameters, and the doublet separation is fixed to tabulated values: $\mu_p = c \ln(\lambda_2^p/\lambda_1^p)$. We fit for known metal doublets in Table 4.2 by summing these functions: $\xi = \xi_{\text{smooth}} + \sum_p \xi_p$. The residuals and their respective fits are in Fig. 4.16. This simple model accurately describes the important features in both correlation functions.

4.4.6 Systematic error budget

We identified four possible sources of systematic errors in our analysis, which are difficult to rectify by models. The continuum is fitted by hand and hard to reproduce. DLA removal is also partly manual in our measurement. For metals, no model at sufficient resolution exists. Our systematic error budget is similar to Chabanier et al. (2019) and based on the fiducial power parameters as given in Section 4.4. We provide each systematic error separately in our files, so that they can be scaled to other values users see fit.

- **Incomplete DLA removal**: Even though we used visual identification and catalogs, we leave room for missed DLAs or other high-density absorbers by assigning 1% incompleteness to this possibility. We multiply the fiducial power estimate with redshift average of the absolute ratio of $P_{\text{kept}}/P_{\text{masked}} - 1$ in Fig. 4.10 and DLA
Continuum systematics affect the largest scales that are outside our conservative range. DLA systematics behave similarly, but they are an order of magnitude smaller. Metal systematics are relevant at low redshifts or high $k$. 

Figure 4.17: Systematic error estimates divided by the SB 1 subtracted bootstrap estimates.
incompleteness ratio.

\[ \sigma_{\text{DLA}} = 0.01 \times P_{\text{fid}}(k) \times \left| \frac{P_{\text{Kept}}}{P_{\text{Masked}}} - 1 \right| \]  \hspace{1cm} (4.27)

We note that the effects of sub-DLAs and LLS that are absent in catalogues can be larger than 1% (Rogers et al., 2018).

- **Continuum:** We assign 10% inefficiency to continuum marginalization and use the redshift average of the absolute ratio of \( P_{\text{NoMar}} / P_{\text{Mar}} - 1 \) in Fig. 4.11. This error is also scaled with the fiducial power.

\[ \sigma_{\text{Cont}} = 0.10 \times P_{\text{fid}}(k) \times \left| \frac{P_{\text{NoMar}}}{P_{\text{Mar}}} - 1 \right| \] \hspace{1cm} (4.28)

Note that continuum errors themselves could be larger than 10%. Our method however prevents these errors from contaminating \( P_{1D} \) measurements by marginalizing out bulk of them, namely a constant and a slope per chunk (not spectrum). Here, our systematic budget really comes from the efficiency of this removal, which we assume to be 90% effective. Furthermore, the modes that are most affected are not in our conservative range.

- **Resolution:** As we have discussed in Section 4.2, seeing conditions alter the resolution. Due to this correction, we expect resolution accuracy to be smaller than normal. XQ-100 reports seeings with an average of 10% precision. SQUAD however is much worse at 40%. We then assign corresponding inaccuracies to \( R \) values.

\[ \sigma_{\text{Res}} = r_m \times P_{\text{fid}}(k) \times 2k^2 \overline{R}^2, \] \hspace{1cm} (4.29)

where \( \overline{R} \) is the average resolution in a given redshift bin and \( r_m \) is 0.1 for XQ-100 and 0.4 for KS. Note that even an undesirable 40% change in resolution for SQUAD-like
spectrum results in only 1–3% change in $P_{1D}$ at $k = 0.1 \text{ s km}^{-1}$. Our reported precision at this value is at least twice as large. Here, we assumed the resolving power is provided with perfect precision by the data sets, and ignored its contribution to the budget.

- **Metal:** There are two reasons to include a systematic error budget for metals: 1) Our statistical estimate of the metal power might be off. 2) Bootstrap error estimates might not reflect the truth. Assuming metal power is nearly constant with $k$ and $z$, we compared the fluctuations between redshifts and bootstrap error estimates. We found on average fluctuations between redshift bins are 6% larger than the bootstrap error estimates. Leaving room for the blending of metal lines we pick 10% for our budget (Day et al., 2019). Using the fiducial power of the side band estimates, the systematic error is given by

$$\sigma_{\text{Metal}}^{(\text{Syst})} = 0.10 \times P_{\text{fid}}^{(\text{SB1})}(k).$$

The results are summarized in Fig. 4.17. DLA and continuum systematics affect the largest scales as expected, but DLA systematics are an order of magnitude smaller. Errors due to resolution inaccuracy becomes relevant near $k = 0.1 \text{ s km}^{-1}$, but they do not overcome statistical errors. Metal systematics are relevant at low redshifts. In our conservative $0.004 \text{ s km}^{-1} < k < 0.1 \text{ s km}^{-1}$ and $z > 1.8$ range, systematic errors are 19% of the statistical errors on average.
4.5 Discussion

4.5.1 Statistical power of our results

We would like to compare the statistical power of our results to the existing measurements. In order to do so, we come up with a crude Fisher forecasting analysis, which replaces the actual $P_{1D}$ measurement with a fiducial power spectrum and only takes the covariance matrix into account. Our model is unquestionably primitive and ignores many complications including thermal state of the IGM. But it will serve as an adequate frame of reference.

Fisher forecast technique is built on the same framework as the quadratic estimator. The function to minimize is

$$\mathcal{L} = -\ln P = \frac{\chi^2}{2} = \frac{1}{2} [d - P(\theta)]^T F [d - P(\theta)], \quad (4.31)$$

where $d$ is the observation ($P_{1D}$ in our case) and $P$ is a model with parameters $\theta$. We limit our simple forecast to a single parameter, the mass of warm dark matter $m_x$. As we are primarily interested in comparing the constraining powers, we ignore different values of $d$ between different measurements and assume a fiducial model as the "truth" for all. Then, the error on $m_x$ is given by

$$\frac{1}{\sigma^2_{m_x}} = \frac{\partial^2 \mathcal{L}}{\partial m_x^2} = P_{m_x}^T F P_{m_x}, \quad (4.32)$$

where $P_{m_x}$ is analytically calculated and the Fisher matrix $F$ is taken from actual measurements. Hence, we need an analytical model describing the response of $P_{1D}$ to $m_x$.

We start building a simple model with the fitting function provided in Bode et al. (2001)
for the transfer function of a warm dark matter only universe:

$$T^2(k) \equiv \frac{P_{\text{WDM}}(k)}{P_{\text{CDM}}(k)} = [1 + (\alpha k)^{2\nu}]^{-\frac{10}{\nu}},$$  \hspace{1cm} (4.33)

where we adopt $\nu = 1.12$ and

$$\alpha = 0.049 \left( \frac{m_x}{1 \text{keV}} \right)^{-1.11} \left( \frac{\Omega_x}{0.25} \right)^{0.11} \left( \frac{h}{0.7} \right)^{1.22} \text{Mpc} h^{-1}$$  \hspace{1cm} (4.34)

from Viel et al. (2005). Current constraints indicate $\alpha < 0.03 \text{Mpc} h^{-1}$ (Murgia et al., 2018), $m_x > 2 \text{keV}$ (Viel et al., 2005) and $m_x > 3.5 \text{keV}$ (Iršič et al., 2017a). We pick $m_x = 4 \text{keV}$ as our fiducial dark matter mass.

We keep the modelling of Ly $\alpha$ forest $P_{3D}$ out of scope of this work, but note that $P_{1D}$ is highly sensitive to small scale physics including Jeans smoothing, reionization history (Hui & Gnedin, 1997; Gnedin & Hui, 1998) and redshift space distortions. Instead, we return to our fiducial fit in equation (4.4). The effect of warm dark matter on $P_{1D}$ can be derived as follows:

$$P_{1D}^{\text{WDM}}(k; m_x) = \int_k^\infty \frac{q dq}{2\pi} P_{3D}(q) T^2(k; m_x)$$  \hspace{1cm} (4.35)

$$P_{3D}(k) = -\frac{2\pi}{k} \frac{dP_{\text{fit}}}{dk}$$  \hspace{1cm} (4.36)

Integration by parts yields

$$P_{1D}^{\text{WDM}}(k; m_x) = P_{\text{fit}}(k) T^2(k; m_x) + \int_k^\infty P_{\text{fit}}(q) dT^2(q)$$  \hspace{1cm} (4.37)

$$P_{1D}^{\text{WDM}}(k; m_x) = P_{\text{fit}}(k) T^2(k; m_x) - 20 \int_k^\infty \frac{P_{\text{fit}}(q) T^2(q; m_x)}{1 + (\alpha q)^{2\nu}} d\ln q.$$  \hspace{1cm} (4.38)

Finally, we use Planck 2018 as our fiducial cosmology (Planck Collaboration et al., 2020b), and convert from distance units in Mpc $h^{-1}$ to velocity units in Ly $\alpha$ forest using
$k_{Ly\alpha} = k_{\text{Mpc}}(1 + z)/H(z)$, which roughly corresponds to a factor of 0.01. Fig. 4.18 shows the suppression in $P_{1D}$ for different warm dark matter masses.

For Iršič et al. (2017c), we add systematic errors in quadrature to the diagonal of the covariance matrix and use all $P_{1D}$ points. This yields $\sigma_{mx} = 2.07$ keV. We note that Iršič et al. (2017a) analysis incorporates some HIRES quasars in addition to XQ-100 only Iršič et al. (2017c) results. The same process yields $\sigma_{mx} = 1.44$ keV for Chabanier et al. (2019). We use metal subtracted result from Walther et al. (2017) and limit to $k < 0.1$ s km$^{-1}$ range. They do not provide further systematic error budget, so the covariance matrix is not modified. This yields $\sigma_{mx} = 1.08$ keV. These numbers are also listed in table 4.3.

For the statistical power of our measurement, we limit ourselves to a conservative $0.004$ s km$^{-1} < k < 0.1$ s km$^{-1}$ and $z > 1.8$ range. This analysis with only statistical errors from bootstrap yields $\sigma_{mx} = 0.36$ keV for our results. We then add our systematic error budget in quadrature to the diagonal of the covariance matrix, which increases the final error on $m_x$ to $\sigma_{mx} = 0.45$ keV. Adding the largest scales $k < 0.004$ s km$^{-1}$ does not improve this number as expected. However, $z = 1.8$ bin brings this value down to 0.30 keV.
Table 4.3: Comparison of different data sets’ statistical warm dark matter mass constraining power in terms of estimated errors, $\sigma_{m_x}$, for fiducial $m_x = 4 \text{ keV}$. We note that Iršič et al. (2017a) analysis incorporates some HIRES quasars in addition to XQ-100 measurement from Iršič et al. (2017c). Without systematic error budget, our results can improve warm dark matter mass constraints by a factor of 3. Additional error budget decreases this factor to 2.4.

To summarize, we expect our results to improve warm dark matter constraints by a factor of 2.4.

Interestingly, including XQ-100 at the chunk level actually yields higher $\sigma_{m_x} = 0.44 \text{ keV}$ with only statistical errors from bootstrap. This increased error from the bootstrap realizations reflects the resolution disparity between data sets.

### 4.5.2 Remaining challenges

We assumed the noise is uncorrelated throughout, even though QMLE is capable of including pixel level correlations. Correlated noise has two effects. First, it changes the weighting in QMLE. This is not a big effect, since these weights roughly stay constant across $k$ bins. A second effect is the uncorrected spurious power. This is removed through the side band power subtraction, though not exactly because we combine data from different instruments.

Furthermore, if the noise is correlated, it should be correlated at the pixel level, which corresponds to very small scales that are outside our conservative range. For example, $k = 0.1 \text{ s km}^{-1}$ corresponds to approximately 20 pixels for KODIAQ and SQUAD. Finally, our error estimates are from bootstraps, which will account for the correlated noise in the error bars. Therefore, we expect our $P_{1D}$ estimates not to be biased due to correlated noise. However, the impact of correlated noise should be revisited as the precision of
$P_{1D}$ increases in the future.

One shortcoming of high-resolution quasar observations is that they are targeted for specific studies of e.g. strong absorbers or metallicity distribution. Murphy et al. (2019) notes that some UVES quasars were specifically targeted due to known DLAs. The original goal of the KODIAQ survey was to study $\text{O}
\text{vi}$ absorption at $z > 2.2$ (O’Meara et al., 2017). Even though we mask or subtract contaminants, their cross correlations with the IGM remain. Therefore, unbiased high-resolution quasar observations are crucial to remedy the effect of this sampling bias.

The continuum errors remain a big challenge for $P_{1D}$ measurements at large scales in general, but especially for these types of quasar spectra. Our marginalization limits the shape of the errors to a constant and a slope, and is not fully descriptive given hand-fitted continua. One could imagine marginalizing out higher order terms and trying to find convergence. However, it is possible that this recipe is not convergent and extra degrees of freedom will eventually wipe out all information. We defer a dedicated analysis to future work. Another option would be finding and marginalizing out dominant terms in continuum by a PCA analysis. Although such templates exist, they come with non-negligible errors with a range of 3–30% (Suzuki et al., 2005). In short, accurately describing the quasar continuum remains a much-needed but difficult task.

4.6 Summary

High-resolution, high-S/N quasar spectra can measure 1D Ly $\alpha$ forest power spectrum at smaller scales compared to large-scale structure surveys. At these scales, $P_{1D}$ is able to constrain the thermal state of the IGM and different dark matter models.

We applied the optimal quadratic estimator to the largest such data set by combining two publicly available data releases (KODIAQ + SQUAD) at the spectrum level, and performed a separate analysis for another publicly available data set XQ-100. We then presented the
weighted average of these two as our final \( P_{1D} \) measurement, and found that it agreed with previous studies with reduced error bars. We identified four systematic error sources for our analysis: incomplete DLA removal, inefficient continuum marginalization, resolution errors and metals. As an advantage of the quadratic estimator, our method is not biased due to gaps and hence free from the respective systematic error. These four systematics are scale and redshift dependent, but smaller than the statistical errors.

Finally, to demonstrate the constraining power of our measurement, we performed a crude, single-parameter Fisher forecast analysis for the warm dark matter mass. We estimate that our results can improve its constraints by a factor of 2.4.
Chapter 5

Future Applications

5.1 Applying the optimal estimator to DESI

A natural next step for my research would be applying the optimal estimator to DESI survey validation (SV) and then to 1-yr data. DESI will obtain spectra for tens of millions of galaxies and quasars, and hundreds of thousands of Ly$\alpha$ quasars in 5 years. With approximately 100 000 quasars, DESI 1-yr data is almost on a par with eBOSS. A precise $P_{1D}$ estimate can tighten neutrino mass constraints, but requires understanding noise properties, spectrograph resolution and continuum errors. The optimal estimator will be an effective tool for DESI’s efforts as it can weigh pixels by noise, marginalize out continuum errors and naturally integrate the full resolution matrix. As opposed to FFT-based estimators, it can also handle long chunks of spectrum (i.e. the redshift evolution within), giving access to the largest scales, and has the additional advantage of not requiring a stringent S/N cutoff, which would allow us to keep a larger portion of the data. Because of these reasons, I believe an early application to DESI can prove valuable to the future of precision cosmology.
5.1.1 Implementing the resolution matrix

Current analyses approximate the effect of spectrograph resolution as wavelength independent and calculate $P_{1D}(k) = \langle P_{\text{raw}}(k)/W_R^2(k) \rangle$, where $W_R(k)$ is the spectrograph window function. However, in general this resolution can be described by a matrix $R$ such that the signal $s$ is multiplied with this resolution matrix in pixel space (Bolton & Schlegel, 2010). This can be naturally expressed as $\tilde{S} = RSR^T$ in optimal estimator formalism. This resolution matrix will be provided in DESI’s pipeline, and its full account would secure accuracy near $k \sim 0.01 \, \text{s km}^{-1}$.

5.1.2 Test and improve continuum marginalization

In Ly$\alpha$ forest analysis, flux $f(\lambda)$ is divided by the quasar continuum $C(\lambda)$ and the mean normalized flux $\bar{F}(\lambda)$ to obtain flux fluctuations $\delta_F(\lambda)$. The errors in this process propagate to mostly large scales and are called the continuum errors $\eta(\lambda)$. These errors are commonly described by a constant shift and a slope, $\eta(\lambda) = a + b \ln \lambda$, which I showed can be marginalized out in the exact limit. Testing this feature with realistic mocks will indicate how well this approximation holds. Furthermore, using a generic piecewise function (i.e. $\eta(\lambda) = \sum_{i=1}^{N} \eta_i(\lambda)$, where $N$ is the number of regions) could be more robust and enable us to measure the largest scales.

5.1.3 Preliminary results from Survey Validation

DESI Survey Validation (SV) is a collection of observations made in a phase prior to the main survey. These observations will be used to test the quality of data, optimize target selection, finalize the survey design and assess data reduction performance. SV approximately started in December 2020 and has produced four internal releases as of June 2021. It also produced some Ly$\alpha$ quasars for initial $P_{1D}$ estimates. These quasar spectra
Figure 5.1: Preliminary DESI SV results. Pipeline noise is currently underestimated in SV. The approximations made in resolution also results in overcorrections at $k > 0.1 \text{ s km}^{-1}$.
are coadded, continuum fit and reduced by DESI’s common Ly \( \alpha \) package PICCA\(^1\). This preliminary analysis has nearly 3500 quasars.

Fig. 5.1 shows preliminary results from SV. FFT estimates are not corrected for noise or resolution, and here for a rough cross check. I note that these results are not treated for DLAs and metals, so disagreements at large and small scales are to be expected. However, one known issue in DESI’s pipeline is the underestimated noise, which results in overall extra power on all scales. This will be fixed in future reductions. Another problem visible in QMLE results is the resolution over-corrections at \( k > 0.01 \) s km\(^{-1}\). This could be due to pipeline reduction problems, but possibly is due to rounding errors that go into approximating the resolving power to keep the number of lookup tables manageable. Its long-term solution is the implementation of full resolution matrix. Over all, these results look promising for the DESI main survey’s future.

### 5.1.4 Realistic DESI mocks

Results from DESI, as well as novel features such as resolution matrix and continuum marginalization in \( P_{1D} \) estimation, will have to be validated and proved on synthetic spectra. The accurate simulations of the entire pipeline are needed to reliably perform these tests. The mocks I generated for this thesis ignore many astrophysical, instrumental and pipeline complications. More realistic simulations incorporate quasar continuum and Poisson noise on observed flux, treat different spectrograph arms with non-Gaussian resolution and carefully add the astrophysical effects of metals and DLAs among others. DESI collaboration developed a code, temporarily named QUICKQUASARS, to simulate the instrument and most of these effects.

In a simple summary, QUICKQUASARS takes transmission files, \( F(\lambda) \), in the absence of noise and resolution effects on a finely spaced wavelength grid. It multiplies these files

\(^1\)I would like to thank Michael Walther for helping me access these data.
Figure 5.2: A QUICKQUASARS simulated spectrum for a quasar at $z = 2.4$ using transmission files generated from this thesis. Ly $\alpha$ emission line falls onto blue (B) arm of the spectrograph at $\lambda \approx 4100$ Å. Similarly, Si IV and C IV emissions fall onto the same arm approximately at 4700 Å and 5200 Å.
with continuum templates, adds noise based on desired exposure time, simulates three arms separately and resamples onto the common DESI wavelength grid of 0.8 Å sized linear pixels. Mocks I generated in this thesis are statistically well-understood at both large and small scales limit, and can be used as input transmission files for QUICKQUASARS. Fig. 5.2 shows one simulated spectrum of a quasar at \( z = 2.4 \). QUICKQUASARS separates spectrum into three spectral arms. Ly\( \alpha \) peak is near 4100 Å and falls onto the blue arm.

### 5.2 Tidal reconstruction for biased tracers

Most of the neutral hydrogen is shielded in galaxies; and the intensity of the 21-cm line is a biased tracer of the underlying dark matter field. Furthermore, if the foreground wedge problem persists, the line-of-sight modes can be more effective in reconstruction as my thesis suggests. To provide optimal efficiency for H\( \text{I} \), tidal reconstruction should be tested on halos of N-body simulations, and modified to account for bias and redshift space distortion parameters if necessary. Moreover, the tidal reconstruction can be extend to galaxy surveys as a result of this project—similar quadratic reconstruction techniques have already caught attention in different frameworks (Li et al., 2020b; Li et al., 2020a; Darwish et al., 2020).

#### 5.2.1 Photometric surveys

Contrary to 21-cm intensity maps, these surveys lose high \( k_z \) modes to photometric redshift errors. In other words, they have high angular resolution, but low line-of-sight resolution. Large-scale angular modes are also contaminated by the survey window function. Tidal reconstruction can be used to estimate the density field at these large angular scales, as it mostly utilizes small-scale angular modes at low \( k_z \). Furthermore, it can also provide a separate power spectrum estimate with different systematics, which originates from three- and four-point functions of small scales.
5.2.2 Using spherical coordinates

Inspired by the weak lensing, tidal reconstruction works on the plane perpendicular to the line of sight, and estimates $\gamma_1 = \partial_x^2 \Phi - \partial_y^2 \Phi$ and $\gamma_2 = 2\partial_x \partial_y \Phi$ at constant $z$ slices, where $\Phi$ is the gravitational potential. This definition is robust in redshift space, but results in utilizing mostly low $k_z$-high $k_\perp$ modes. When foregrounds leak into angular modes due to instrumental imperfections in 21-cm maps (foreground wedge), bulk of these modes get washed out, and the remaining modes form a cone around the line of sight, i.e. $\theta < \theta_{\text{wedge}}$ in spherical coordinates. Therefore, to efficiently use low $l$-high $k$ modes, it is better to estimate the shears at constant $\theta$ slices. Although more complex, this framework would most effectively utilize the accessible modes of 21-cm intensity maps.
Appendix A

Tidal Reconstruction

A.1 Time Dependent Functions

The redshift dependent function $f$ is given by

$$f(k, z) = -2D_{1st}(z) + F(z) \left(2 - \frac{d \ln P_l(k, z)}{d \ln k}\right), \quad (A.1)$$

where $D(z)$ is the linear growth function and

$$F(z) = \int_z^{\infty} \frac{D(z'')dz''}{H(z'')} \int_z^{z''} \frac{(1 + z') dz'}{H(z')} \quad (A.2)$$

$$D_{1st}(z) = \int_z^{\infty} \frac{Y(z')dz'}{H(z')^2} \left[\frac{H(z)}{D(z)} D(z') - H(z')\right]. \quad (A.3)$$

We defined the following intermediate functions to simplify expressions,

$$T(z) = D(z)(1 + z) \quad (A.4)$$

$$W(z) = H(z)D'(z) - H'(z)D(z) \quad (A.5)$$

$$Y(z) = T(z)D(z)/W(z). \quad (A.6)$$
A.2 Details on Reconstruction Theory

Three dimensional convergence field is given by the integration in equation (2.47). We start by writing

\[ \tilde{T}_{\gamma_i} \]

\[ \tilde{T}_{\gamma_i}(k', k - k') = 2k^2 x^2 + k^2 y^2 \times \left[ (k^2 x - k^2 y)\tilde{T}_{\gamma_{i}}(k', k - k') + 2k_x k_y \tilde{T}_{\gamma_{i}}(k', k - k') \right] \]

(A.7)

Since \( T_{\gamma_i} \) is symmetric, \( \tilde{T}_{\gamma_i}(k', k - k') \) is symmetric as well.

We can reshape this expression into the following:

\[ \tilde{T}_{\gamma_i}(q, k - q) = -\frac{2k^2 w(q)w(k - q)}{3k^4_{\perp}} \times \left[ \frac{k^2_{\perp} q^2_{\perp} - 2(k_{\perp} \cdot q_{\perp})^2 + (k_{\perp} \cdot q_{\perp})k^2_{\perp}}{q|k - q|} \right]. \]

(A.9)

We want to evaluate \( \tilde{T}_{\gamma_i}(q, k - q) \) in spherical coordinates. We start by fixing \( k_{\perp} \) and assigning an angle \( \phi \) between \( k_{\perp} \) and \( q_{\perp} \). We continue using \( p \equiv k - q \) and define \( t \equiv \cos \phi \). Then, the following relations hold:

\[ \mathbf{k} \cdot \mathbf{q} = k_{\perp} q_{\perp} t + k_z q_z \]

(A.10)

\[ p_{\perp} = \sqrt{k^2_{\perp} + q^2_{\perp} - 2k_{\perp} q_{\perp} t} \]

(A.11)

\[ p_z = k_z - q_z \]

(A.12)

\[ p = \sqrt{k^2 + q^2 - 2\mathbf{k} \cdot \mathbf{q}}. \]

(A.13)

We further define \( q_z \equiv q \cos \theta = qv \). Then \( q_{\perp} = q\sqrt{1 - v^2} \). We can rewrite the relations
above in terms of these new spherical coordinates.

\[ \mathbf{k} \cdot \mathbf{q} = k_\perp q_t \sqrt{1 - v^2} + k_z q_v \]  \hspace{1cm} (A.14)

\[ p_\perp = \sqrt{k_\perp^2 + q^2(1 - v^2) - 2k_\perp q_t \sqrt{1 - v^2}} \]  \hspace{1cm} (A.15)

\[ p_z = k_z - q_v \]  \hspace{1cm} (A.16)

Finally, \( \tilde{T}_{\kappa3D} \) can be expressed in terms of the same coordinates.

\[ \tilde{T}_{\kappa3D}(q, v, t, k_\perp, k_z) = -\frac{2 k_\perp^2}{3k_\perp^2} w(q_\perp, q_z) w(p_\perp, p_z) \]  \hspace{1cm} \times \left[ q(1 - v^2)(1 - 2t^2) + k_\perp t \sqrt{1 - v^2} \right] \hspace{1cm} (A.17)

\[ \delta G(k) = \delta_R(k) + \frac{1}{2} \int \frac{d^3q}{(2\pi)^3} \delta_R(q) \delta_R(k - q) + \Theta(\delta^3) \]  \hspace{1cm} (A.18)

\[ = S_R(k) \delta(k) - \frac{1}{2} \int \frac{d^3q}{(2\pi)^3} S_R(q) S_R(p) \delta(q) \delta(p) \]  \hspace{1cm} (A.19)

\[ = S_R(k) \left( \delta_l(k) + \int \frac{d^3q}{(2\pi)^3} F_2(q,p) \delta_l(q) \delta_l(p) \right) \]  \hspace{1cm} (A.20)

\[ - \frac{1}{2} \int \frac{d^3q}{(2\pi)^3} S_R(q) S_R(p) \delta_l(q) \delta_l(p) \]  \hspace{1cm} (A.21)

where we have modified the kernel to include smoothing.

\[ F_{2,G,R}(k_1, k_2) = F_2(k_1, k_2) - \frac{1}{2} \frac{S_R(k_1)S_R(k_2)}{S_R(|k_1 + k_2|)} \]  \hspace{1cm} (A.22)

There are two modifications to previous modified bispectrum \( B_T \) expression. First, we
multiply $B_T$ by $S_R(q)S_R(p)$. Second, we substitute $F_{2,G} \rightarrow F_{2,G,R}$.

$$B_{T,R} = S_R(q)S_R(p)2[F_2(q,p)P_l(q)P_l(p)$$

$$+ F_{2,G,R}(p,-k)P_l(p)P_l(k) + F_{2,G,R}(q,-k)P_l(q)P_l(k)]$$

$$B_{T,R}(q,p) = S_R(q)S_R(p) \times \left[ B - \frac{S_R(p)S_R(k)}{S_R(q)}P_l(p)P_l(k) - \frac{S_R(q)S_R(k)}{S_R(p)}P_l(q)P_l(k) \right]$$

(A.23)
Appendix B

Lyman-alpha Forest

B.1 Saddle Point Approximation for the Mean Flux of Log-normal Mocks

We can find a closed analytic form for the mean flux in equation 3.24 using saddle point approximation assuming the integrand is large around its maximum. We start by finding the derivatives of the function in the exponential:

\[ \phi(\delta) \equiv -\frac{\delta^2}{2\sigma^2} - xe^{-2a\delta}, \]  
\[ \phi'(\delta) = -\frac{\delta}{\sigma^2} - 2axe^{-2a\delta}, \]  
\[ \phi''(\delta) = -\frac{1}{\sigma^2} - 4a^2xe^{-2a\delta}, \]

where the \( z \) dependence of \( x(z) \) and \( a(z) \) are suppressed for clarity. Solving for the maximum \( \phi'(\delta_*) = 0 \) yields \( -\delta_* = 2a\sigma^2xe^{2a\delta_*} \), for which the solution is given by the Lambert \( W \) function:

\[ \delta_*(z) = -\frac{1}{2a} W\left(4a^2\sigma^2x\right) \]
Note that $\delta_* < 0$. Substituting this solution back into $\phi$ and $\phi''$ yields:

$$\phi(\delta_*) = \frac{\delta_*}{2a\sigma^2}(1 - a\delta_*), \quad (B.5)$$

$$\phi''(\delta_*) = -\frac{1 - 2a\delta_*}{\sigma^2}. \quad (B.6)$$

Now we can approximate the integration around the maximum.

$$\bar{F}(z) \approx \frac{1}{\sigma\sqrt{2\pi}} \int_{\delta_* - \epsilon}^{\delta_* + \epsilon} \exp \left[ \phi(\delta_*) + \frac{(\delta - \delta_*)^2}{2\phi''(\delta_*)} \right] d\delta \quad (B.7)$$

$$\approx \frac{e^{\phi(\delta_*)}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp \left[ \phi''(\delta_*) \frac{\delta^2}{2} \right] d\delta = \frac{e^{\phi(\delta_*)}}{\sqrt{-\phi''(\delta_*)\sigma^2}}, \quad (B.8)$$

and note that $-\phi''(\delta_*)\sigma^2 = 1 - 2a(z)\delta_*(z)$. We can further define $d_*(z) = a(z)\delta_*(z)$ and $\sigma(z) = a(z)\sigma$ to simplify this expression.

$$\bar{F}(z) \approx [1 - 2d_*(z)]^{-1/2} \exp \left\{ \frac{d_*(z)[1 - d_*(z)]}{2\sigma^2(z)} \right\} \quad (B.9)$$

As noted in the main section, this approximation starts to deviate from the truth at $z \lesssim 3$ up to 4%. It would be interesting to use this function as a fitting template for the real data.
Bibliography


Akitsu K., Takada M., 2018, Phys. Rev. D, 97, 063527 16


Bandura K., et al., 2014. International Society for Optics and Photonics, p. 914522, doi:10.1117/12.2054950 7, 10, 14

Barkana R., Loeb A., 2007, Reports on Progress in Physics, 70, 627 8


DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036 9, 12, 45, 61, 75


Iršič V., et al., 2017a, Phys. Rev. D, 96, 023522 111, 112, 113


Matsubara T., 2008a, Phys. Rev. D, 77, 063530 5


McDonald P., et al., 2006, ApJS, 163, 80 3, 12, 13, 44, 45, 50, 53, 74, 82, 89, 100

McQuinn M., 2016, ARA&A, 54, 313 6, 7


135


