Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **473,** 3019–3038 (2018)
Advance Access publication 2017 September 28

doi:10.1093/mnras/stx2525

The SDSS-DR12 large-scale cross-correlation of damped Lyman alpha systems with the Lyman alpha forest

Ignasi Pérez-Ràfols,[1,2]★ Andreu Font-Ribera,[3] Jordi Miralda-Escudé,[1,4] Michael Blomqvist,[5] Simeon Bird,[6] Nicolás Busca,[7] Hélion du Mas des Bourboux,[8] Lluís Mas-Ribas,[9] Pasquier Noterdaeme,[10] Patrick Petitjean,[10] James Rich[8] and Donald P. Schneider[11,12]

[1]*Institut de Ciències del Cosmos, Universitat de Barcelona/IEEC, Barcelona, E-08028, Catalonia*
[2]*Departament de Física Quàntica i Astrofísica, Universitat de Barcelona/IEEC, Barcelona E-08028, Catalonia*
[3]*Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT, UK*
[4]*Institució Catalana de Recerca i Estudis Avançats, Barcelona E-08028, Catalonia*
[5]*Aix Marseille Univ, CNRS, LAM, Laboratoire d'Astrophysique de Marseille, 13013 Marseille, France*
[6]*Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*
[7]*APC, Université Paris Diderot-Paris 7, CNRS/IN2P3, CEA, Observatoire de Paris, 10, rue A. Domon & L. Duquet, 75013 Paris, France*
[8]*IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France*
[9]*Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029 Blindern, NO-0315 Oslo, Norway*
[10]*Universiteé Paris 6 et CNRS, Institut d'Astrophysique de Paris, 98bis blvd. Arago, F-75014 Paris, France*
[11]*Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, PA 16802, USA*
[12]*Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA*

Accepted 2017 September 25. Received 2017 September 25; in original form 2017 August 7

ABSTRACT

We present a measurement of the damped Ly α absorber (DLA) mean bias from the cross-correlation of DLAs and the Ly α forest, updating earlier results of Font-Ribera et al. (2012) with the final Baryon Oscillations Spectroscopic Survey data release and an improved method to address continuum fitting corrections. Our cross-correlation is well fitted by linear theory with the standard ΛCDM model, with a DLA bias of $b_{\rm DLA} = 1.99 \pm 0.11$; a more conservative analysis, which removes DLA in the Ly β forest and uses only the cross-correlation at $r > 10\,h^{-1}$ Mpc, yields $b_{\rm DLA} = 2.00 \pm 0.19$. This assumes the cosmological model from Planck Collaboration (2016) and the Ly α forest bias factors of Bautista et al. (2017) and includes only statistical errors obtained from bootstrap analysis. The main systematic errors arise from possible impurities and selection effects in the DLA catalogue and from uncertainties in the determination of the Ly α forest bias factors and a correction for effects of high column density absorbers. We find no dependence of the DLA bias on column density or redshift. The measured bias value corresponds to a host halo mass $\sim 4 \times 10^{11}\,h^{-1}\,{\rm M}_\odot$ if all DLAs were hosted in haloes of a similar mass. In a realistic model where host haloes over a broad mass range have a DLA cross-section $\Sigma(M_h) \propto M_h^\alpha$ down to $M_h > M_{\rm min} = 10^{8.5}\,h^{-1}\,{\rm M}_\odot$, we find that $\alpha > 1$ is required to have $b_{\rm DLA} > 1.7$, implying a steeper relation or higher value of $M_{\rm min}$ than is generally predicted in numerical simulations of galaxy formation.

Key words: galaxies: intergalactic medium – cosmology: cosmological parameters – cosmology: observations – cosmology: large-scale structure of the Universe.

1 INTRODUCTION

Damped Ly α absorbers (DLAs) are absorption systems of high neutral hydrogen column density, usually defined as $N_{\rm H\,I} \geq 2 \times 10^{20}\,{\rm cm}^{-2}$ (Wolfe et al. 1986). At these column densities, the damped profile of the hydrogen Ly α line is measurable even in low-resolution spectra and with the superposition of the Ly α forest, allowing the column density to be directly measured from the absorption profile. This lower limit on $N_{\rm H\,I}$ is also related (depending on the ionization parameter, or ratio of the gas density to the photoionization rate) to absorption systems which, owing to

★ E-mail: iprafols@icc.ub.edu

self-shielding of the external cosmic ionizing background radiation, have most of their hydrogen in atomic form (e.g. Vladilo et al. 2001). For reviews on DLAs, see e.g. Wolfe, Gawiser & Prochaska (2005) and Barnes, Garel & Kacprzak (2014).

DLAs are therefore a probe to any gaseous systems that have condensed to high enough densities to become self-shielding, which are naturally associated with sites of galaxy formation. In the standard cold dark matter (CDM) model of structure formation, we expect these sites to be located in haloes over a broad range of mass, from those of dwarf galaxies to groups of massive galaxies. Measurements of the incidence rate and column density distribution imply a contribution to the matter density of the atomic gas contained in these systems of $\Omega_{\rm DLA} \simeq 10^{-3}$ at redshifts $2 < z < 3.5$ (Péroux et al. 2003; Prochaska, Herbert-Fort & Wolfe 2005; Noterdaeme et al. 2009; Prochaska & Wolfe 2009; Noterdaeme et al. 2012; Zafar et al. 2013; Crighton et al. 2015; Padmanabhan, Choudhury & Refregier 2016). This accounts for ∼2 per cent of all baryons in the Universe, which is comparable to the fraction of baryons in stars at the same redshifts. These absorption systems are therefore regarded as reservoirs of atomic gas clouds for the formation of the stellar component of galaxies, and they are crucial to understand how galaxies can be gradually formed from gas that is accreted in galactic haloes.

The study of metal absorption lines associated with DLAs is a powerful tool to study the dynamics and evolution of this gas and has revealed that DLAs typically have low metallicities distributed over a broad range of $10^{-3}\,{\rm Z}_{\odot}$ to $1\,{\rm Z}_{\odot}$, and on average decreasing gradually with redshift (Kulkarni & Fall 2002; Vladilo 2002; Prochaska et al. 2003; Kulkarni et al. 2005; Rafelski et al. 2012; Jorgenson, Murphy & Thompson 2013; Møller et al. 2013; Neeleman et al. 2013; Mas-Ribas et al. 2017).

This implies that the gas reservoir in DLAs has been enriched from material ejected by stars, which were formed either in low-mass galaxies that later merged into the DLA host halo together with the gas, or in a galaxy in the DLA host halo itself. Absorption lines from low and high-ionization species associated with DLAs suggest a broad range of densities and temperatures (Wolfe & Prochaska 2000; Prochaska & Wolfe 2002; Fox et al. 2007a,b). The kinematics of these low and high ionization gas phases differ, and a complex structure of absorption components at different velocities are often seen in high spectral resolution data, reflecting a clumpy structure with typical velocity ranges of $\sim 100\,{\rm km\,s^{-1}}$ (Prochaska & Wolfe 1997, 1998; Wolfe & Prochaska 1998). Several models of gaseous galactic haloes have been proposed to account for these observations (see e.g. Haehnelt, Steinmetz & Rauch 1998; McDonald & Miralda-Escudé 1999; Fumagalli et al. 2011; Cen 2012; Rahmati & Schaye 2014; Bird et al. 2015; Neeleman, Prochaska & Wolfe 2015).

Despite this rich information on the velocity structure of DLAs, the mass distribution of their host halo masses is not well known. One way to characterize this distribution is to analyse the clustering properties of DLAs. In the limit of large scales, where linear theory holds, the correlation function of any population of objects that trace the primordial mass perturbations is equal to the mass autocorrelation times the square of the bias factor (e.g. Cole & Kaiser 1989; Mo & White 1996). In redshift space, where all our observations are done, the same relation holds adding a redshift space distortion term (Kaiser 1987). The bias factor of haloes increases with their mass in a way that can be accurately predicted both analytically (see e.g. Sheth & Tormen 1999) and from sophisticated numerical simulations (see e.g. Tinker et al. 2010). Therefore, if every DLA is associated with a dark matter halo, a measurement of the mean bias factor of any population of DLAs tells us the mean bias factor of their host haloes and constrains in a powerful way their mass distribution.

A first method for measuring the DLA bias, $b_{\rm DLA}$, is by measuring the DLA autocorrelation. This approach, however, requires a large sample and has not been attempted so far due to smaller number of DLAs compared to quasars. A more convenient method is to use the cross-correlation with another tracer population. The first cross-correlation that was detected was with Lyman break galaxies in the vicinity of the quasar lines of sight (Cooke et al. 2006), but owing to their small sample size (only 11 DLAs), the bias factor could only be constrained to $1.3 < b_{\rm DLA} < 4$.

The Baryon Oscillations Spectroscopic Survey (BOSS; Dawson et al. 2013) in the Sloan Digital Sky Survey III (SDSS-III; Eisenstein et al. 2011) allowed for a very large sample of quasars and DLAs to be obtained, which opened the way for measuring a variety of cross-correlations on scales much larger than had been attainable before. The other tracer of cosmological density fluctuations that is most useful for obtaining the DLA bias factor turns out to be the Ly α forest absorption, because of its presence in every quasar spectrum over a broad redshift range. The cross-correlation with the Ly α forest was first measured by Font-Ribera et al. (2012, hereafter FR12) using the data release (DR9) of BOSS, with a sample of 7458 DLAs, and a value $b_{\rm DLA} = 2.17 \pm 0.20$ was obtained, where the error reflects only uncertainties from the observational determination of the cross-correlation, and not from the model used to derive the bias. The main modelling uncertainty lies in the bias and redshift distortion parameter of the Ly α forest, because only the product of bias factors of the two tracer populations can be determined. In that work, the first determination of the Ly α forest bias factors by Slosar et al. (2011) was used. This measurement was based on the early data release of the BOSS sample of quasar spectra containing the Ly α forest.

This paper is an update to the measurement of the cross-correlation of DLAs and the Ly α forest by FR12. Using the entire DR12 sample, we can decrease the errorbars of this measurement and we can better explore the dependence of the bias factor on the DLA column density and the redshift evolution. A dependence of the mean bias factor on any DLA properties can provide powerful constraints on galaxy formation models and tests on the predictive accuracy of cosmological numerical simulations (e.g. Bird et al. 2014). In addition, we use the improved estimate of the Ly α forest bias factors by Bautista et al. (2017), implying a substantial reduction of our systematic errors in deriving the DLA bias as well.

We start by describing the data sets used to derive the DLA bias in Section 2. An improved estimator for the cross-correlation is described in Section 3. Section 4 explains the model used to fit the DLA bias. Then, in Section 5 we present our results. A detailed comparison with previous measurements and a study of the model dependences of the DLA bias measurement is made in Section 6. Finally, the cosmological implications for the halo masses hosting DLAs are discussed in Section 7, and we summarize our conclusions in Section 8. Throughout this paper, we use a flat ΛCDM cosmology, with $\Omega_m = 0.3156$, $\Omega_b = 0.0492$, $h = 0.6727$, $n_s = 0.9645$ and $\sigma_8 = 0.831$, as reported by Planck Collaboration (2016).

2 DATA SAMPLE

In this section, we describe the data sets used in this study, based on the DR12 of SDSS-III (Gunn et al. 1998; York et al. 2000; Gunn et al. 2006; Eisenstein et al. 2011; Bolton et al. 2012; Smee et al. 2013), which is the final data release of BOSS (Dawson

et al. 2013). The quasar target selection used in BOSS is summarized in Ross et al. (2012), and combines different targeting methods described in Yèche et al. (2010), Kirkpatrick et al. (2011) and Bovy et al. (2011).

We measure the cross-correlation of two tracers of the underlying density field: the number density of DLAs and the Ly α absorption along a set of lines of sight. The DLAs used as tracers are designated here as *DLA sample* and the quasar lines of sight, where the Ly α absorption is measured are designated as *Ly α sample*. All the quasars used to find the DLAs and measure the Ly α absorption spectra are in the DR12Q catalogue (Pâris et al. 2017).

2.1 DLA sample

For the DLA sample, we use an early version of the DR12 extension of the DLA catalogue from Noterdaeme et al. (2012). This sample contains a total of 34 050 DLAs candidates with column density $N_{\rm H\,I} \geq 10^{20}\,{\rm cm}^{-2}$. For convenience, from here on we will refer to these DLAs candidates simply as DLAs. We note that the precise number of DLAs varies slightly with the different versions of the catalogue that were produced, but the inclusion or exclusion of the small number of objects that differ amongst the versions does not affect in any significant way the results in this paper. Although the strict definition of a DLA requires its column density to be above $2 \times 10^{20}\,{\rm cm}^{-2}$, systems with column density down to $10^{20}\,{\rm cm}^{-2}$ are still identified with high efficiency in BOSS data and are not expected to sharply change their nature. We will test the dependence of the properties of DLAs we measure with column density. Out of the 34 050 DLAs, there are 12 which have the catalogue identifier *ThingID* set to -1, which indicates an error in the pipeline data reduction for these objects. They are excluded from the final sample.

We now describe several cuts we apply to the remaining 34 038 DLAs to obtain our DLA sample with an increased purity compared to that of the catalogue. Purity of our sample is important because objects included in the catalogue that are not real or are at the wrong redshift will decrease the measured bias of DLAs, while confusion with other types of absorption systems (e.g. Lyman limit systems with extra Ly α forest absorption around them in high noise spectra) might increase the measured bias if these absorption systems have a higher bias than DLAs. On the other hand, completeness is less important: eliminating a fraction of the real DLAs will only result in an increase of the errors of the cross-correlation without modifying it systematically, as long as the probability of inclusion of the DLAs in the catalogue is not correlated with its large-scale cosmological environment. The cuts applied here are the same as those in FR12, except that we add additional ones to obtain different samples and test the dependence of our results on them, and they are as follows.

First cut: DLA redshift, $z_{\rm DLA}$. We include only DLAs in the redshift range of $2.0 \leq z_{\rm DLA} < 3.5$. Outside this redshift interval, DLAs have few nearby lines of sight with sufficient signal-to-noise ratio in the Ly α forest to be useful to measure the correlation, and we eliminate them to have a well-defined redshift interval. This reduces our sample to 31 059 DLAs.

Second cut: continuum-to-noise ratio (CNR) ≥ 3. The CNR of the Ly α forest spectral region, defined in Noterdaeme et al. (2012), provides a good estimate of the data quality over the region of interest and is independent of the presence of DLAs. Since it is more difficult to detect DLAs in noisier spectra, we apply this second cut to increase the purity of the sample without drastically reducing the number of systems. A total of 23 568 DLAs survive this cut.

Table 1. Summary of DLA samples A, C1 and C2, and subsamples of sample A with the indicated redshift and column density bins.

Name	Description	Number of DLAs
A	Full DLA sample	13 734
C1	Full DLA sample excluding cuts 4–6	23 342
C2	Full DLA sample excluding cut 6	19 655
Z1	DLAs with $z_{\rm DLA} < 2.25$	3348
Z2	DLAs with $2.25 \leq z_{\rm DLA} < 2.5$	3455
Z3	DLAs with $2.5 \leq z_{\rm DLA}$	6931
N1	DLAs with $\log\left(N_{\rm H\,I}/{\rm cm}^{-2}\right) < 20.26$	4448
N2	DLAs with $20.26 \leq \log\left(N_{\rm H\,I}/{\rm cm}^{-2}\right) < 20.63$	4683
N3	DLAs with $20.63 \leq \log\left(N_{\rm H\,I}/{\rm cm}^{-2}\right)$	4603

Third cut: eliminating broad absorption line (BAL) systems . The systems can produce wide O VI absorption with profiles that can be confused with the Voigt profiles of DLAs. We exclude all the DLAs found in the spectra of quasars with any positive Balnicity Index, as listed in the DR12Q catalogue, leaving 23 342 DLAs.

Fourth cut: DLAs close to the Ly α emission line. All systems within a velocity separation of $v_c < 5000\,{\rm km\,s}^{-1}$ from the quasar redshift are eliminated. This condition is equivalent to requiring $\lambda_r \geq 1195.39$ Å, where λ_r is the quasar rest-frame wavelength at which the DLA absorption line is centred. This reduces our sample to 21 408 DLAs.

Fifth cut: DLAs close to the O VI emission line. An excess of DLAs with 1005 Å $< \lambda_r <$ 1037 Å was found in FR12, likely caused by BAL contamination. Removing all DLAs with λ_r in this interval reduces our sample to 19 655 DLAs.

Sixth cut: DLAs in the Ly β forest. All the systems bluewards of the Ly β emission line are removed. This is done because, as found in Mas-Ribas et al. (2017), a small fraction of the DLAs detected bluewards of the Ly β line are in fact Ly β absorption features for which the Ly α line is not properly identified, and are then confused with the Ly α line of a DLA with the method of Noterdaeme et al. (2009). This cut causes a considerable further reduction of our sample to 13 734 DLAs.

The final sample contains a total of 13734 DLAs. We emphasize again that the purity of the sample is more important than its completeness. However, we understand that the fourth, fifth and especially the sixth cut exclude an important amount of DLAs, most of which will be true DLAs. To analyse the importance of these cuts in the final measurement, different DLA samples are studied in this work. We label the final sample considering all cuts as data set A, and the final sample considering only the cuts that are most useful to remove contaminants (i.e. not applying cuts four to six) as C1. Finally, we label data set C2 to be the sample resulting from the application of all constraints save the sixth. In this sample, the same cuts as in FR12 are applied, allowing for a more direct comparison. The properties of the three data sets are summarized in Table 1.

We separate data set A in bins of the DLA redshift and column density. The bins are chosen in data set A to obtain subsamples with similar signal-to-noise ratio in the measured cross-correlation. We label the redshift subsamples Z1, Z2 and Z3, and the column density subsamples N1, N2 and N3, with properties listed in Table 1. Fig. 1 shows the distribution of the total DLA sample in redshift and column density. The bins used to define the subsamples Z1 to Z3 and N1 to N3 are indicated as red solid lines and are given in Table 1.

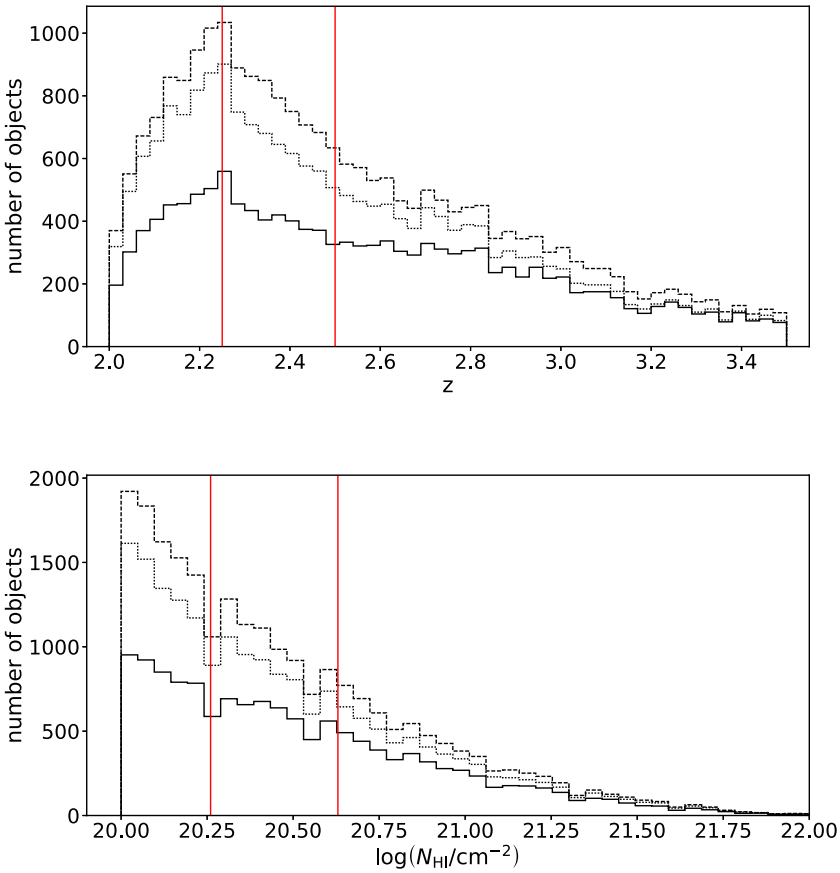

Figure 1. Distribution of the DLAs in samples A (solid), C1 (dashed) and C2 (dotted) in redshift (top panel) and column density (bottom panel). Solid red lines show the bins used to construct the subsamples (see Table 1).

2.2 Ly α Sample

For the Ly α sample, we use the same set of Ly α spectra of DR12 as in Busca et al. (2013), with a total of 157 922 spectra containing over 27 million Ly α pixels. We use their *analysis pixels* that are the average of every three pixels of the actual co-added spectra, because our cross-correlation measurements do not depend on small-scale variations and this saves computational time. Throughout the rest of this paper, *pixel* refers to analysis pixels unless otherwise stated. The effective width of these pixels is $(\Delta\lambda/\lambda)c = 207\,\mathrm{km\,s^{-1}}$.

The Ly α transmission fluctuation at every pixel i with wavelength λ_i and measured flux f_i is defined as

$$\delta_i = \frac{f_i}{C_q(\lambda_i)\,\overline{F}(z_i)} - 1. \quad (1)$$

Here, $C_q(\lambda)$ is the quasar continuum (or unabsorbed flux) and $\overline{F}(z)$ is the mean transmitted fraction at the Ly α absorber redshift. The pixel redshift is $z_i = \lambda_i/\lambda_{\mathrm{Ly}\,\alpha} - 1$. We use the quasar continuum designated as *method 1* in Busca et al. (2013), which assumes a universal shape of the quasar rest-frame continua except for a multiplicative factor that is linear in wavelength that allows for a variable slope of the quasar continuum that is fitted to the Ly α forest region. We refer the reader to Busca et al. (2013) for a more detailed description of this method.

An important difference relative to FR12 is that we correct the Ly α forest transmission for the DLAs that are identified in the DR12 DLA catalogue that we use. If this is not done, the DLA-Ly α cross-correlation includes a component that is caused by the DLA autocorrelation, owing to the contribution from DLAs to the Ly α absorption spectra. We apply this correction in the same way as Bautista et al. (2017): for every DLA in the catalogue, we compute its absorption Voigt profile and we eliminate any pixels in which the computed DLA transmission is less than 0.8. We then correct all other pixels in the spectrum by dividing the measured transmission by the computed DLA transmission. This eliminates only the detected DLAs, so the absorption of any undetected systems remains (these systems are generally of low column density in low signal-to-noise spectra). We discuss how this is modelled in Section 4.

3 CROSS-CORRELATION

3.1 Estimator for the cross-correlation

In this section, we describe the method used to compute the cross-correlation of DLAs and the Ly α transmission fluctuation δ_i, and its covariance matrix. The method is similar to that used in FR12, although the broad-band uncertainties arising from the continuum fitting of quasar spectra are treated in a different way. FR12 used a simple estimator of the cross-correlation ξ as a function of the parallel and perpendicular components of the separation vector $\boldsymbol{r}$ between a DLA and a Ly α pixel, given by

$$\tilde{\xi}^A = \frac{\sum_{i\in A} w_i \delta_i}{\sum_{i\in A} w_i}, \quad (2)$$

where the sum is overall DLAs and overall the pixels i located within a bin A of the separation $\boldsymbol{r}$ from a DLA, and the weights w_i are defined to optimize the accuracy of the measurement of ξ^A. They then performed a *mean transmission correction* (MTC) to compensate for the effects of the quasar continuum fitting. Note that a given pixel i of the Ly α forest appears as many times in the sum in equation (2), as there are DLAs at a separation from the pixel within bin A.

We adopt a different approach, following the one used in Bautista et al. (2017). We present a brief description of the method here, and a more extended and detailed explanation in Appendix B. The goal is to remove from the cross-correlation the part that is strongly affected by systematics related to the continuum fit, by using an adequate projector. The effect of this projector can then be taken into account in the modelling, eliminating the need for the MTC. Our assumption is that the measured Ly α transmission fluctuation $\delta^{(m)}$ differs from the true one, $\delta^{(t)}$, by a linear additive function,

$$\delta_i^{(m)} = \delta_i^{(t)} + a + b \log \lambda_i , \tag{3}$$

where a and b are unknown for each forest. That is, we assume that a linear approximation to the continuum in the region of the Ly α forest adequately describes the effect of all the systematic calibration errors in the observed spectrum and of having fitted a continuum to it. Although in this paper, we use this linear expansion in $\log \lambda_i$, this method works the same way if the linear fit is assumed in λ_i instead. We define a projector P_f for each forest f that removes this unknown part by subtracting a weighted linear regression to the forest, so that the projected measured and true fluctuations are equal:

$$\delta_i \equiv \sum_{j \in f} P_{f,ij} \delta_j^{(m)} = \sum_{j \in f} P_{f,ij} \delta_j^{(t)} . \tag{4}$$

The sums are overall pixels j that belong to the same forest f as pixel i. From this point on, we use δ_i to mean the projected transmission fluctuation, after subtracting the weighted linear regression by applying the projector P_f. A more detailed derivation of the equation for this projector is given in Appendix B. The cross-correlation in bin A is then expressed by exactly the same equation (2), except that now δ_i is understood to have been projected.

In general, this projector can introduce an artificial non-vanishing correlation at large scales, arising from a mean value of $\delta = P_f \delta^{(m)}$ at a given redshift that is not equal to zero, because only the mean value of $\delta^{(m)}$ in narrow redshift bins was initially required to be zero. We solve this by computing the cross-correlation of the mean transmission value, $\bar{\delta}_i$, at the redshift z_i of pixel i, designated as $\tilde{\xi}^A_{\rm sky}$, using the same equation (2), and then subtracting it as a correction. The final cross-correlation is

$$\xi^A = \tilde{\xi}^A - \tilde{\xi}^A_{\rm sky} . \tag{5}$$

For the cross-correlation between DLAs and the Ly α forest, this correction is negligible at our current level of precision, but this needs not be the case in general.

3.2 Covariance matrix

The covariance of the cross-correlation at two bins A and B is equal to

$$\begin{aligned} C^{AB} &\equiv \left\langle \xi^A \xi^B \right\rangle - \left\langle \xi^A \right\rangle \left\langle \xi^B \right\rangle \\ &= \frac{1}{S^{AB}} \sum_{i \in A} \sum_{j \in B} w_i w_j \, \zeta_{ij} , \end{aligned} \tag{6}$$

where ζ_{ij} is the Ly α forest autocorrelation of the values of δ at pixels i and j, and each of the two sums are again understood to be overall Ly α forest pixels and all the DLAs at separations within the bins A or B. The normalization factor is

$$S^{AB} = \sum_{i \in A} \sum_{j \in B} w_i w_j . \tag{7}$$

As discussed in FR12, there are three main contributions to the correlation ζ_{ij}. First, there is a noise component that we assume to be uncorrelated amongst different pixels, and is therefore present only for $i = j$. This contribution arises from the fact the same Ly α pixel contributes several times to the evaluation of ξ at different bins when it is paired with different DLAs. Secondly, there is a contribution produced by continuum fitting errors inducing a correlation amongst pixels in the same forest. Finally, different Ly α pixels are intrinsically correlated due to the physical Ly α forest autocorrelation. This entire autocorrelation ζ_{ij} can be measured directly from the data, but in practice it is computationally expensive to compute the covariance matrix taking into account the correlation amongst pixels in different forests out to a large transverse separation, because of the large number of DLA-Ly α pixels pairs-of-pairs involved in the sum of equation (6). In this work, we neglect the contribution to the covariance matrix of pixels in different forests. We find, however, that it is important to measure the change of ζ with redshift. Once we restrict this autocorrelation to pixel pairs on a single forest, ζ_{ij} can be expressed as a function of the redshift z and the separation $n = j - i$ in number of pixels between j and i along the line of sight,

$$\zeta(z, n) = \frac{\sum_{i, z_i = z} w_i w_{i+n} \delta_i \delta_{i+n}}{\sum_{i, z_i = z} w_i w_{i+n}} , \tag{8}$$

where the sum is overall pixels i that have redshift $z_i = z$, and the δ_i are as usual the projected transmission fluctuations.

We compute this autocorrelation in redshift bins of width $\Delta z = 0.0037$ for n up to 5. We have checked that further increasing the maximum value of n does not modify the recovered covariance matrix, while it increases the computational time.

3.3 Distortion matrix

Having applied the projection to the data to eliminate the most important continuum fit systematics, we need to correct the model we fit to include the effect of this projection. The mixing of the δ variables in the same forest due to this projection implies that the projected cross-correlation in bin A, ξ^A_p, is related to the model cross-correlation ξ_m by a distortion matrix D,

$$\xi^A_p = \sum_B D^{AB} \xi^B_m . \tag{9}$$

The distortion matrix element D^{AB} relates the projected cross-correlation in bin A to the model cross-correlation at all bins B, and can be directly computed from the quasar positions in the survey and the redshift range of each forest being used, with the same method that was used in Bautista et al. (2017). The resulting ξ_p is the one that is compared to the projected measured cross-correlation ξ in equation (5) to fit any given model. The detailed way we compute the distortion matrix is explained in Appendix B (Section B3).

3.4 Bootstrap errors

The errors obtained when computing the covariance matrix rely on the validity of the approximations we have made. One of the most

important approximations is that we include DLA-Ly α forest pairs of pairs only when the two Ly α pixels are in the same forest. We also neglect errors associated with spectral calibration, which are difficult to model reliably for including them in a direct calculation of the covariance matrix. It is therefore important to test the validity of our errors by computing them alternatively using the bootstrap method.

We divide the survey into subsamples using the plate number of the observations. Each DLA-Ly α pair is always assigned to the plate that the Ly α pixel belongs to. Using the 2400 regions defined by the plates, a total of 100 bootstrap samples are generated. We compute the cross-correlation for each of these bootstrap samples and then we fit our model (see Section 4), modified by the distortion matrix mentioned above and using the covariance matrix to compute the χ^2. The distortion and covariance matrices are computed for the whole sample and not modified for each of the 100 bootstrap resamplings. The bootstrap errors of model parameters are computed in the standard way, equal to the dispersion of the best-fitting parameter values obtained in the bootstrap samples.

4 FITTING THE CROSS-CORRELATION

This section describes the linear theory model that is used to fit the measured DLA-Ly α forest cross-correlation. All the actual fits are computed with the publicly available fitting code BAOFIT (Kirkby et al. 2013, see http://darkmatter.ps.uci.edu/wiki/DeepZot/Baofit).

In the limit of large scales, linear theory predicts the form of the cross-correlation of any two tracers of the large-scale mass-density fluctuations. The limit of large scales is broadly expected to apply when the relative mass-density fluctuation is small compared to unity at the redshift of our observations, but the precision at which linear theory is reliable depends on the tracer. For the Ly α forest, linear theory often works surprisingly well because the transformation from optical depth to transmission fraction suppresses the contribution from highly overdense regions, which develop the largest non-linearities, to the measured correlations.

In real space, any biased tracer should have the same linear fluctuations as the mass-density, except for a linear biased factor. For example, the Ly α forest transmission fluctuation at any pixel i, after being smoothed three dimensionally over a large scale, would simply be related to the mass fluctuation δ_m smoothed in the same way by $\delta_i = b_{\mathrm{Ly}\,\alpha}\delta_m$, if the effects of peculiar velocities were somehow eliminated. In Fourier space, the same relation holds for the Fourier modes. However, observations can only be done in redshift space, where peculiar velocity gradients enhance the amplitude of each Fourier mode according to the expression found by Kaiser (1987),

$$\delta_i = b_{\mathrm{Ly}\,\alpha}\left(1+\beta_{\mathrm{Ly}\,\alpha}\mu_k^2\right)\,\delta_m, \tag{10}$$

where $\beta_{\mathrm{Ly}\,\alpha}$ is the redshift distortion parameter, and μ_k is the cosine of the angle between the Fourier mode vector and the line of sight. The density fluctuations of DLAs also have their own bias and redshift distortion parameter, and the linear cross-power spectrum of the two types of objects is equal to

$$P_{\mathrm{DLA,Ly}\alpha}\left(\boldsymbol{k},z\right) = b_{\mathrm{DLA}}\left(1+\beta_{\mathrm{DLA}}\mu_k^2\right) \times b_{\mathrm{Ly}\,\alpha}\left(1+\beta_{\mathrm{Ly}\alpha}\mu_k^2\right)\,P_L(k,z)\,G(\boldsymbol{k})\,S(k_\parallel), \tag{11}$$

where $P_L(k, z)$ is the linear matter power spectrum. We have introduced also two smoothing functions of the cross-correlation, which are multiplicative functions in Fourier space: $S(k_\parallel)$ accounts for the spectrograph resolution and binning of the Ly α forest spectra, and $G(\boldsymbol{k})$ accounts for the binning used to compute the cross-correlation function. For the calculation of the linear power spectrum P_L, BAOFIT uses templates that were computed for our specific cosmology using CAMB at the reference redshift $z_{\mathrm{ref}} = 2.3$ (Kirkby et al. 2013).

The DLA-Ly α cross-power in equation (11) depends only on the product of the two bias factors b_{DLA} and $b_{\mathrm{Ly}\alpha}$. We can therefore infer the value of one of the bias factors only if the other one, as well as the normalization of P_L, is independently constrained. The two redshift distortion bias factors have effects that are also difficult to separate, and only one of them can be measured in practice from the shape of the cross-correlation in redshift space. Previous analyses of the BOSS Collaboration (see e.g. Blomqvist et al. 2015; Delubac et al. 2015; Bautista et al. 2017) have studied in detail the Ly α forest autocorrelation and obtained constraints on the Ly α forest bias factors. We use the values listed in table 3 of Bautista et al. (2017): $\beta_\alpha = 1.663 \pm 0.085$ and $b_\alpha(1+\beta_\alpha) = -0.325 \pm 0.004$, at a reference redshift $z_{\mathrm{ref}} = 2.3$. We fix these two Ly α forest parameters to their mean values from this measurement. The errors and modelling uncertainties of the Ly α forest bias factors obtained in this way introduce systematic errors in our derived DLA bias factor, which are discussed in detail in Section 6.6.

We do not include in our model any additive broad-band function to measure the form of the cross-correlation, which can arise from spectral calibration systematics and continuum fitting in the Ly α forest region, and have been used in previous studies of the BOSS Ly α data where the focus was in measuring the narrow-band feature of the Baryon Acoustic Oscillation peak in the correlation function (e.g. Font-Ribera et al. 2014; Blomqvist et al. 2015; Delubac et al. 2015; Bautista et al. 2017).

The model is evaluated at the mean values of the parallel and perpendicular components of the separation vector, $r_\parallel$ and $r_\perp$, for each of the bins of the measured cross-correlation, and at the mean redshift of our sample. For the evolution with redshift, we assume that $b_\alpha \propto (1+z)^{2.9}$, and that b_{DLA} and the redshift distortion parameters β_α and β_{DLA} are constant. This evolution of b_α follows that measured from previous Ly α autocorrelation studies (e.g. McDonald et al. 2006), and we shall see below that a constant b_{DLA} with redshift is consistent with our results. Including the linear growth factor, this implies that the amplitude of the cross-power spectrum in equation (11) evolves approximately as $(1+z)^{0.9}$. We fix $\beta_{\mathrm{DLA}} b_{\mathrm{DLA}} = f(\Omega) = 0.968897$, assuming that there is no peculiar velocity gradient bias for DLAs.

The term $G(\boldsymbol{k}) = G_\parallel\left(k_\parallel\right)\,G_\perp(k_\perp)$ corrects for the binning in $r_\parallel$ and $r_\perp$ that averages the cross-correlation over a bin. We use $G_\parallel(k_\parallel) = \mathrm{sinc}^2(\Delta_\parallel k_\parallel/2)$ and $G_\perp(k_\perp) = \mathrm{sinc}^2(\Delta_\perp k_\perp/2)$, as in Bautista et al. (2017), where $\Delta_\parallel$ and $\Delta_\perp$ are the bin sizes. In this work, they are both equal to $2\,h^{-1}\,\mathrm{Mpc}$. We also correct for the spectrometer resolution and for the averaging of the three spectrometer pixels into analysis pixels, by approximating the convolution of a Gaussian and a top-hat as a new Gaussian, $S\left(k_\parallel\right) = \exp[-k_\parallel^2/(2\sigma_S^2)]$. The contribution to the variance σ_S from the point spread function (PSF) of the BOSS instrument is set to $\sigma_{\mathrm{PSF}} = 0.61\,h^{-1}\,\mathrm{Mpc}$ (in comoving units), which corresponds to a full-width half-maximum $R = (\Delta\lambda/\lambda)^{-1} = 2000$ at the reference redshift $z_{\mathrm{ref}} = 2.3$. The averaging of three spectrometer pixels, which have a top-hat full width $\Delta\lambda/\lambda = 3\ln(10)\times 10^{-4} = 6.91\times 10^{-4}$ (Busca et al. 2013), contributes an additional dispersion $\sigma_p = cH^{-1}(z)(1+z)\,\Delta\lambda/\lambda/\sqrt{12} = 0.57\,h^{-1}\,\mathrm{Mpc}$, also in comoving units. The overall dispersion is $\sigma_S = \sqrt{\sigma_{\mathrm{PSF}}^2+\sigma_p^2} = 0.83\,h^{-1}\,\mathrm{Mpc}$.

In Section 2, we explained how the absorption profiles of DLAs also contribute to the Ly α forest transmission and therefore to the measured DLA-Ly α cross-correlation. While the detected DLAs

are corrected, many DLAs remain undetected in low signal-to-noise spectra, and all the absorption systems with column densities $N_{\rm H\,I} < 10^{20}\,{\rm cm}^{-2}$, which are not considered to be DLAs but also have damped absorption wings, contribute to the Ly α transmission. These systems cannot be removed or corrected directly in the data, and therefore their effect needs to be corrected from the measured cross-correlation.

In general, the measured cross-correlation is the sum of the cross-correlations of DLAs with several populations of objects that contribute to the absorption in the Ly α forest spectra. The population of hydrogen absorbers including unidentified DLAs and systems of lower column density that have significant damped wings is designated as high-column density systems, or high-column-density systems (HCDs). In addition to these, some metal lines with wavelengths close to the Ly α line can also contribute significantly to the cross-correlation, and were modelled in Bautista et al. (2017). We ignore these metal lines here, because the signal-to-noise ratio of the DLA-Ly α cross-correlation is smaller than the Ly α autocorrelation, and the effect of metal lines is not clearly discernible in our results; this is further addressed in Section 6.7. We include only the HCDs as an additive contamination,

$$\xi^A_{\rm obs} = \xi^A_{\rm DLA-Ly\alpha} + \xi^A_{\rm DLA-HCD}, \tag{12}$$

where $\xi^A_{\rm DLA-Ly\alpha}$ is the Fourier transform of the power spectrum in equation (11). The cross-correlation with HCDs is assumed to be the Fourier transform of the same linear theory form of the cross-power as our model for DLAs:

$$P_{\rm DLA-HCD} = b_{\rm DLA}\, b_{\rm HCD}\, (1+\beta_{\rm DLA}\mu_k^2) \times (1+\beta_{\rm HCD}\mu_k^2) P_L(k,z) F_{\rm HCD}(k_\parallel), \tag{13}$$

where the function $F_{\rm HCD}(k_\parallel)$ is introduced to approximately model the average wavelength profile of HCDs, and is set to

$$F_{\rm HCD}\left(k_\parallel\right) = \sin(L_{\rm HCD}k_\parallel)/(L_{\rm HCD}k_\parallel), \tag{14}$$

where $L_{\rm HCD}$ is a parameter that reflects the width of the absorption wings of HCDs. We use the values found in Bautista et al. (2017) to fit the observed Ly α autocorrelation, listed in their table 3: $b_{\rm HCD} = -0.0288$, $\beta_{\rm HCD} = 0.681$ and $L_{\rm HCD} = 24.34\,h^{-1}$ Mpc. The bias $b_{\rm HCD}$ is assumed to evolve with redshift in the same way as the Ly α forest bias, $b_{\rm HCD} \propto (1+z)^{2.9}$, for reasons of computational efficiency (this evolution makes very little difference to the computed effect of HCDs; the value given above is at the reference redshift $z_{\rm ref} = 2.3$), and the other two parameters are assumed to be independent of redshift.

5 RESULTS

We have measured the cross-correlation for all the samples listed in Table 1, with bin sizes $\Delta_\parallel = \Delta_\perp = 2\,h^{-1}$ Mpc, out to a maximum separation of $80\,h^{-1}$ Mpc in both the parallel and perpendicular directions. In this section, all the model parameters as described in the previous section are fixed, and we fit only $b_{\rm DLA}$. Note that the DLA redshift distortion parameter, $\beta_{\rm DLA} = f(\Omega)/b_{\rm DLA}$, also varies with $b_{\rm DLA}$; this is, however, a small effect, because the redshift distortions of the cross-correlation are dominated by $\beta_{\rm Ly\alpha}$, and variations of $\beta_{\rm DLA}$ in all the results we present are small. Neglecting the variation of $\beta_{\rm DLA}$, fitting the bias $b_{\rm DLA}$ is equivalent to fitting the amplitude of our cross-correlation model with a fixed shape to the data, and this amplitude is proportional to $b_{\rm DLA}b_{\rm Ly\alpha}\sigma_8^2$, where σ_8^2 is the standard quantity to express the normalization of the power spectrum P_L.

Our cross-correlation model assumes linear theory, and therefore we exclude bins at a small value of $r = (r_\parallel^2 + r_\perp^2)^{1/2}$ in the fits to reduce the impact of non-linearities on our result. For each sample, we perform two fits: a conservative one that excludes bins with $r < r_{\rm min} = 10\,h^{-1}$ Mpc and a more generous one excluding only bins with $r < r_{\rm min} = 5\,h^{-1}$ Mpc. In addition, all fits exclude bins with $r > 90\,h^{-1}$ Mpc to better define the radius range of our measurements; we shall see that the cross-correlation signal is not clearly detected beyond $r \gtrsim 60\,h^{-1}$ Mpc.

5.1 Measured cross-correlation and DLA bias

The measured values of $b_{\rm DLA}$ are summarized in Table 2. Results for the A, C1 and C2 samples (see Section 2) are presented in Section 5.1; the redshift and column density dependences are explored in Section 5.2 using the subsamples Z1 to Z3 and N1 to N3, and the scale dependence of the bias factor is investigated in Section 5.3.

Our fiducial result to which we refer for all comparisons is the fit to sample A, which yields $b_{\rm DLA} = 2.00 \pm 0.19$ for $r_{\rm min} = 10\,h^{-1}$ Mpc, and $b_{\rm DLA} = 2.06 \pm 0.14$ for $r_{\rm min} = 5\,h^{-1}$ Mpc, at a reference redshift $z_{\rm ref} = 2.3$. In general, the fits reported in Table 2 have covariance matrix errors lower than the errors derived by the bootstrap technique. This is likely because the Ly α transmission correlations in different forests are neglected when computing the covariance matrix. The bootstrap errors should therefore be considered as more reliable.

The values of χ^2 of the fit to the measured cross-correlation indicate that our model is fully consistent with the data for sample A, and marginally inconsistent at the $\sim 3-\sigma$ level for sample C1, for both values of $r_{\rm min}$. This may be due to a contamination of the signal introduced by false DLAs that appear near the Ly α and O VI emission lines of quasars, which are not removed in sample C1.

The results of the DLA-Ly α cross-correlation as a function of $r_\parallel$ are shown for various bins of $r_\perp$ in Fig. 2, for sample A. We have rebinned the cross-correlation measurements into wider bins than the ones used for computing the fits in both $r_\parallel$ and $r_\perp$, for display purposes only, recomputing the plotted errors in these wider bins using our covariance matrix. Results are shown only out to $r_\perp = 60\,h^{-1}$ Mpc, even though our measured cross-correlation is used in all the bins out to $r_\perp = 80\,h^{-1}$ Mpc. These results are also shown as a contour plot with smoothed contours in Fig. 3 (left-hand panel).

The black solid and red dashed lines in Fig. 2 are our best-fitting models for $r_{\rm min} = 10\,h^{-1}$ Mpc and $r_{\rm min} = 5\,h^{-1}$ Mpc, respectively. In practice, the two curves are nearly identical (the difference in $b_{\rm DLA}$ is only 3 per cent) and can hardly be distinguished. The curves are not shown in bins at small r that were not used for the fit, although when the model is averaged into the wider bins for plotting purposes, we include all bins even if they are not used in the fit to facilitate a correct comparison with the data points. The model for the case $r_{\rm min} = 5\,h^{-1}$ Mpc is also presented in a contour format in Fig. 3 (right-hand panel).

When relaxing the cuts imposed in the sample A, we find that $b_{\rm DLA}$ is slightly lower in samples C1 and C2. This may be partly due to a decreased purity when we eliminate some of the cuts imposed on sample A, although the differences are consistent with statistical errors.

Table 2. Summary of the fitted $b_{\rm DLA}$ for each DLA subsample, with the values of χ^2 for the fits with only one free parameter. The values of the bias are given at the reference redshift $z_{\rm ref} = 2.3$. Errors are obtained from our computed covariance matrix, and also using the bootstrap method (shown in parenthesis). See Table 1 for the subsample definitions.

	$r_{\rm min} = 10\,h^{-1}$ Mpc		$r_{\rm min} = 5\,h^{-1}$ Mpc	
Data set	$b_{\rm DLA}$	χ^2(d.o.f.)	$b_{\rm DLA}$	χ^2(d.o.f.)
A	2.00 ± 0.15(0.19)	2817.43 (2864-1)	2.06 ± 0.11(0.14)	2854.08 (2896-1)
C1	1.93 ± 0.11(0.13)	3019.44 (2864-1)	1.97 ± 0.08(0.10)	3065.79 (2896-1)
C2	1.97 ± 0.12(0.14)	2911.86 (2864-1)	1.99 ± 0.09(0.11)	2950.26 (2896-1)
Z1	2.40 ± 0.24(0.31)	2906.85 (2864-1)	2.36 ± 0.17(0.21)	2936.71 (2896-1)
Z2	1.39 ± 0.25(0.29)	2875.71 (2864-1)	1.90 ± 0.18(0.21)	2944.79 (2896-1)
Z3	2.27 ± 0.29(0.31)	2807.96 (2864-1)	1.92 ± 0.20(0.23)	2855.82 (2896-1)
N1	2.05 ± 0.26(0.32)	2844.55 (2864-1)	2.09 ± 0.19(0.26)	2869.06 (2896-1)
N2	2.33 ± 0.26(0.32)	2929.53 (2864-1)	2.17 ± 0.18(0.23)	2955.24 (2896-1)
N3	1.60 ± 0.26(0.28)	2,847.15 (2,864-1)	1.92 ± 0.18(0.20)	2891.66 (2896-1)

5.2 Bias dependence on redshift and column density

The left-hand panel in Fig. 4 shows the DLA bias in three redshift bins, derived from the cross-correlations of samples Z1, Z2 and Z3 (see Table 1). The results are shown with solid errorbars, with horizontal ones indicating the redshift range of each subsample, for both values of $r_{\rm min}$, and are also tabulated in Table 2. There is no evidence for any redshift evolution of the DLA bias. For $r_{\rm min} = 5\,h^{-1}$ Mpc, the scatter of the DLA bias in the three redshift bins is a bit larger than expected, but we believe this is attributable to statistical noise (using the bootstrap errors the measured scatter corresponds to a $\sim 2-\sigma$ fluctuation). Results for the larger sample C1 split into three redshift bins, shown with dotted errorbars, give a smaller scatter; these will be presented in more detail later in Section 7.3.

The dependence of the DLA bias on column density, obtained from the subsamples N1, N2 and N3, is shown in the right-hand panel of Fig. 4, with values tabulated in Table 2. Again, there is no evidence for any dependence on $N_{\rm H\,I}$ for either of the two values of $r_{\rm min}$.

5.3 Scale dependence of the bias factor

We now test if our measured cross-correlation agrees with the theoretically expected radial dependence in linear theory of the ΛCDM model for $P_L(k)$. If this model is correct, there should be no radial dependence of $b_{\rm DLA}$, except at small scales where non-linear effects may be important. We repeat the fit of the sample A cross-correlation to our fiducial model restricted to bins in rings in the $(r_\parallel, r_\perp)$ plane, defined by $2^{(i-1)/2} r_{\rm min} < r < 2^{i/2} r_{\rm min}$, with $i = 0, 1, 2, ..., 8$.

The results of these fits are shown in Fig. 5 and Table 3. While there is no clear dependence of the DLA bias on r, and most of the values of χ^2 are consistent with a good fit for all the rings, we note that the χ^2 value for the second ring is particularly bad. The probability of obtaining such a value is about half a per cent. If this was our only measurement, then the bad χ^2 might indicate that the linear model is starting to fail at these small scales, but see Lochhaas et al. (2016) for a more detailed analysis. However, obtaining such a high χ^2 in one out of nine measurements is not as unlikely. This suggests that the linear theory ΛCDM model correctly predicts the cross-correlation we have measured, as expected if DLAs are associated with dark matter haloes that trace the underlying dark matter distribution (e.g. Mo & White 1996). The cyan line in Fig. 5 shows the result of cumulative fits to all $r < r_{\rm min}$, with the grey band, indicating the $1-\sigma$ error. This error increases with $r_{\rm min}$, as the radial range of the fit is reduced.

We note that at small scales, there is no clear variation of the DLA bias from linear theory down to the smallest radii we test, as might be expected from non-linearities. The saturation of absorption lines in the Ly α forest naturally acts as a mask of the contribution from highly overdense regions to cross-correlations, making linear theory predictions surprisingly accurate down to rather small scales. We therefore consider that our results for the DLA bias with $r_{\rm min} > 10\,h^{-1}$ Mpc are not significantly affected by non-linearities in the cross-correlation. The lack of any clearly visible spreading of contours in $r_\parallel$ at small scales in Fig. 3 also shows that the combination of intrinsic velocity dispersions and redshift errors in our sample A of DLAs is small.

6 COMPARISON WITH PREVIOUS RESULTS AND MODEL DEPENDENCE OF THE DLA BIAS

We now analyse in detail the model dependence of our result on the mean DLA bias. To facilitate the comparison with the previous result of FR12, our reference results in this section will be for sample C2 and $r_{\rm min} = 5\,h^{-1}$ Mpc, which is $b_{\rm DLA} = 1.99 \pm 0.09$ from Table 2. FR12 used an equivalent sample for DR9, and the same value of $r_{\rm min}$, and obtained $b_{\rm DLA} = 2.17 \pm 0.20$. However, the difference with our result does not just arise from using DR12 instead of DR9, but from the following differences in the analysis and the model that is fitted to the data:

(i) The bias fitting method.
(ii) The cross-correlation estimator and covariance matrix.
(iii) Correction for the continuum fitting distortions.
(iv) Larger DR12 data set versus DR9.
(v) Cosmological model.
(vi) Ly α forest bias parameters.
(vii) New additions to the cross-correlation model.

In the rest of this section, we start with the original result of FR12 and change these factors one by one to see how each of them affects the result of $b_{\rm DLA}$. The last three points also account for the main model dependence of our result, discussed in Sections 6.5–6.7. To help the reader track all the effects and changes caused on $b_{\rm DLA}$, a list is provided in Table 4.

6.1 Bias fitting method

The fitting of the model to the data was done in FR12 with an MCMC code written especially for that paper. We have used in this

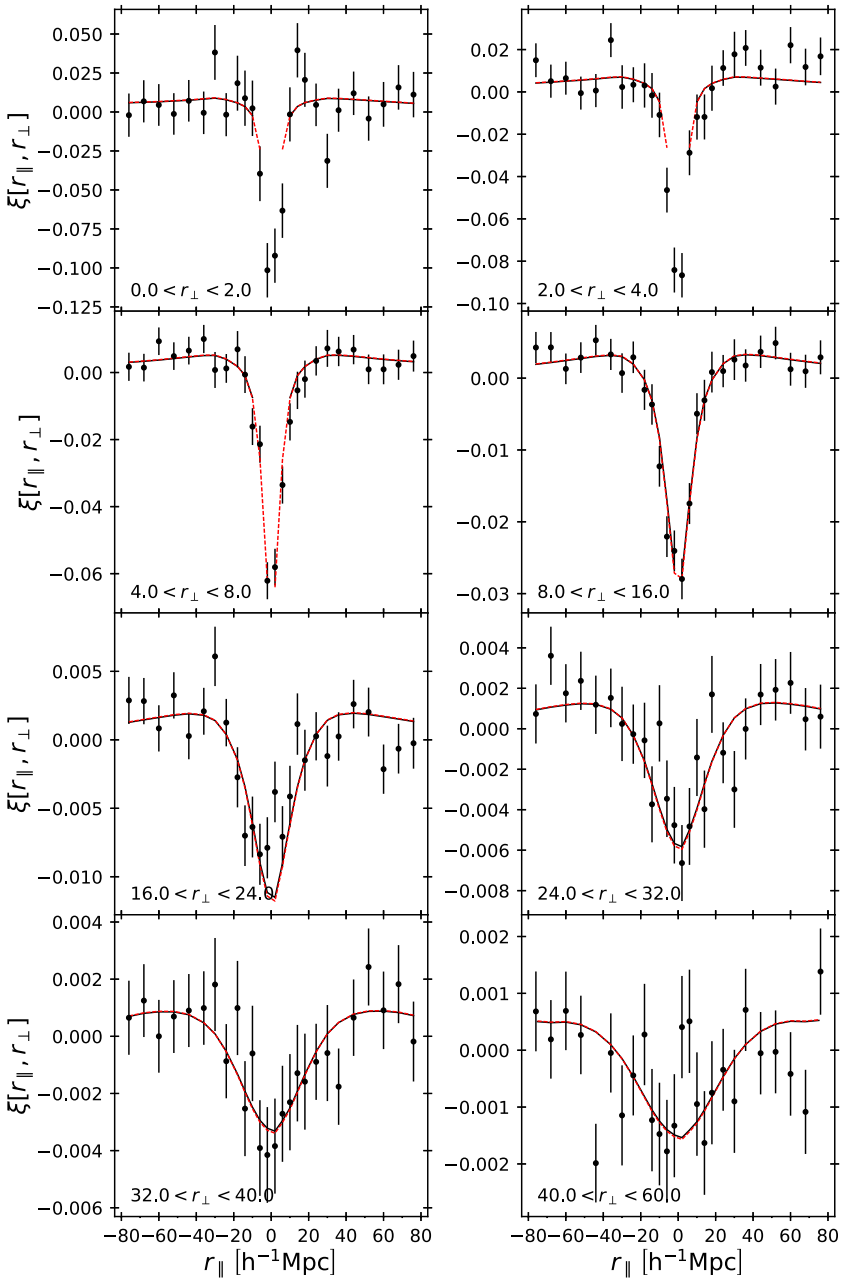

Figure 2. Cross-correlation of DLAs and Ly α forest as a function of $r_\parallel$ for various bins in $r_\perp$, in comoving h^{-1} Mpc, for sample A. Black circles show the measured cross-correlation for sample A. Solid black lines and dashed red lines correspond to the best-fitting model considering $r_{\rm min} = 10\,h^{-1}$ Mpc and $r_{\rm min} = 5\,h^{-1}$ Mpc, respectively, and are nearly equal and hard to distinguish in the figure. Data and models have been rebinned to wider bins than used in the analysis to plot this figure.

paper the BAOFIT code (Kirkby et al. 2013), which computes errors from the second derivatives of the χ^2 function computed from the covariance matrix that is provided. There may therefore be slight differences in the results obtained with the two codes. To test this difference, we have run BAOFIT to fit the $b_{\rm DLA}$ parameter with exactly the same values of the cross-correlation over the same bins, and the same covariance matrix that was computed in FR12 from the DR9 data.

The continuum fitting method of FR12 was different than the one used here, and a correction of the continuum fitting effects called the MTC was applied there to the fitted model, which was quite different from our distortion matrix correction described in Section 3.3. To take out differences in the fitted model, we compare to the FR12 result for their NOCOR case, in which no MTC correction was applied to the model, and the value obtained was $b_{\rm DLA} = 2.00 \pm 0.19$. We use exactly the same cosmological model and Ly α forest bias parameters as were used in FR12, and we eliminate the factors $G(\boldsymbol{k})$ and $S(k_\parallel)$ in equation (11) and our corrections for the distortion matrix and the presence of HCDs to fit to exactly the same cross-correlation model as in FR12. Our result is $b_{\rm DLA} = 2.01 \pm 0.17$.

We therefore conclude that the main effect of the different fitting method is that the errorbars from BAOFIT using the covariance matrix

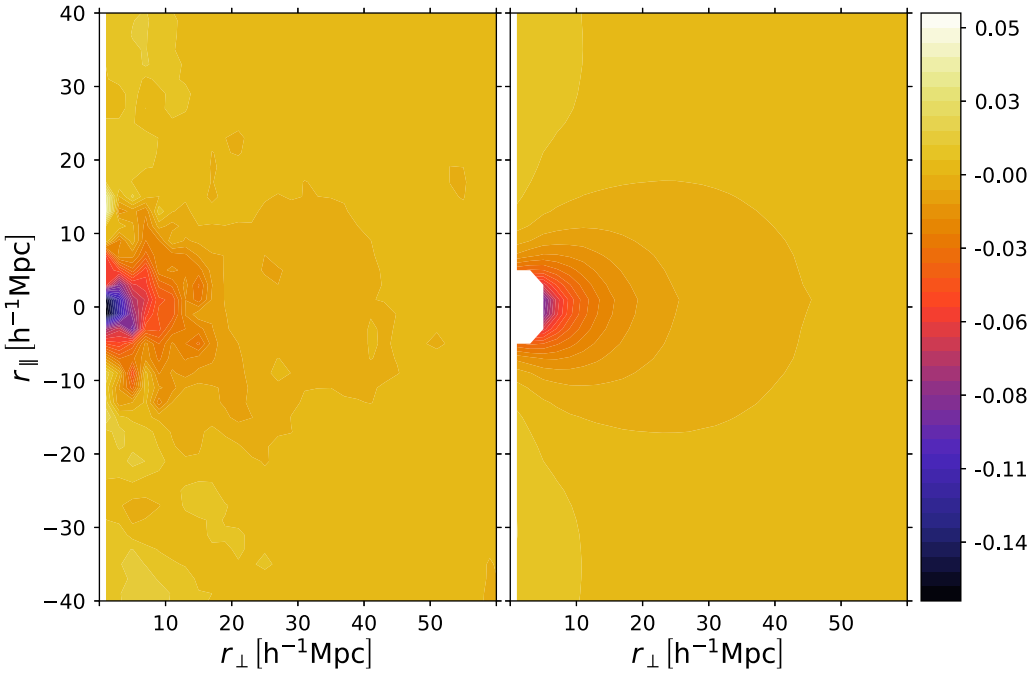

Figure 3. Smoothed contour plots of the measured DLA-Ly α cross-correlation (left, sample A) and best-fitting theoretical model considering bins with $5\,h^{-1}$ Mpc $< r = (r_\parallel^2 + r_\perp^2)^{1/2} < 90\,h^{-1}$ Mpc (right).

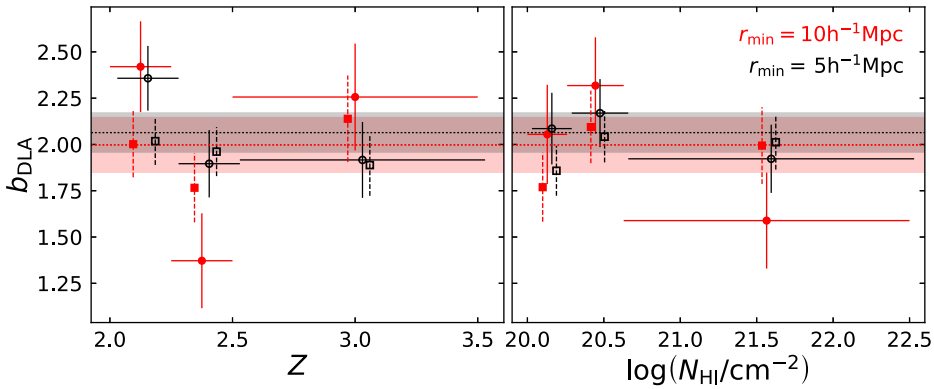

Figure 4. DLA bias versus redshift (left) and log ($N_{\mathrm{H\,I}}$) (right) obtained from subsamples Z1, Z2 and Z3, and N1, N2 and N3, respectively (see Table 1). Black open circles and red closed circles with solid errorbars are fit results for $r_{\mathrm{min}} = 5$ and $10\,h^{-1}$ Mpc, respectively. Dotted lines are the result for sample A, with $1-\sigma$ errors indicated by shaded regions. Squares are equivalent to circles and triangles, but computed from samples ZC1 to ZC3, and NC1 to NC3, described in Section 7.3. The bins in redshift and column density are the same for all cases (shown only for the solid errorbars). Except for red solid circles, points are horizontally shifted to avoid overlap.

Table 3. DLA bias versus r from sample A with the fit restricted to bins with $r \in [r_{\mathrm{min}}, r_{\mathrm{max}})$, in units of h^{-1} Mpc.

r_{min}	r_{max}	b_{DLA}	χ^2 (d.o.f.)
3.54	5.00	2.03 ± 0.30	0.97 (2-1)
5.00	7.07	2.22 ± 0.21	20.19 (8-1)
7.07	10.00	2.18 ± 0.20	16.75 (24-1)
10.00	14.14	2.32 ± 0.25	35.74 (38-1)
14.14	20.00	2.09 ± 0.30	83.19 (80-1)
20.00	28.28	1.44 ± 0.37	163.14 (154-1)
28.28	40.00	2.05 ± 0.44	330.91 (320-1)
40.00	56.57	1.94 ± 0.49	640.28 (620-1)
56.57	80.00	1.59 ± 0.53	1227.02 (1260-1)

are ∼10 per cent smaller than those from the MCMC code used in FR12.

6.2 The cross-correlation estimator and covariance matrix

We now use our own method to determine the continuum of the observed spectra and the values of the Ly α transmission and to estimate the cross-correlation and covariance matrix, using only sample C2 limited to the DR9 data set, i.e. the same data used by FR12. Our Ly α forest data also include the masking and correction for DLAs in the catalogue of Noterdaeme et al. (2014), which were not included in FR12. We fit b_{DLA} exactly as before, using the same cosmological model and Ly α forest bias parameters as FR12, and not including any of the corrections that were not included in FR12. The result we find is $b_{\mathrm{DLA}} = 1.94 \pm 0.15$.

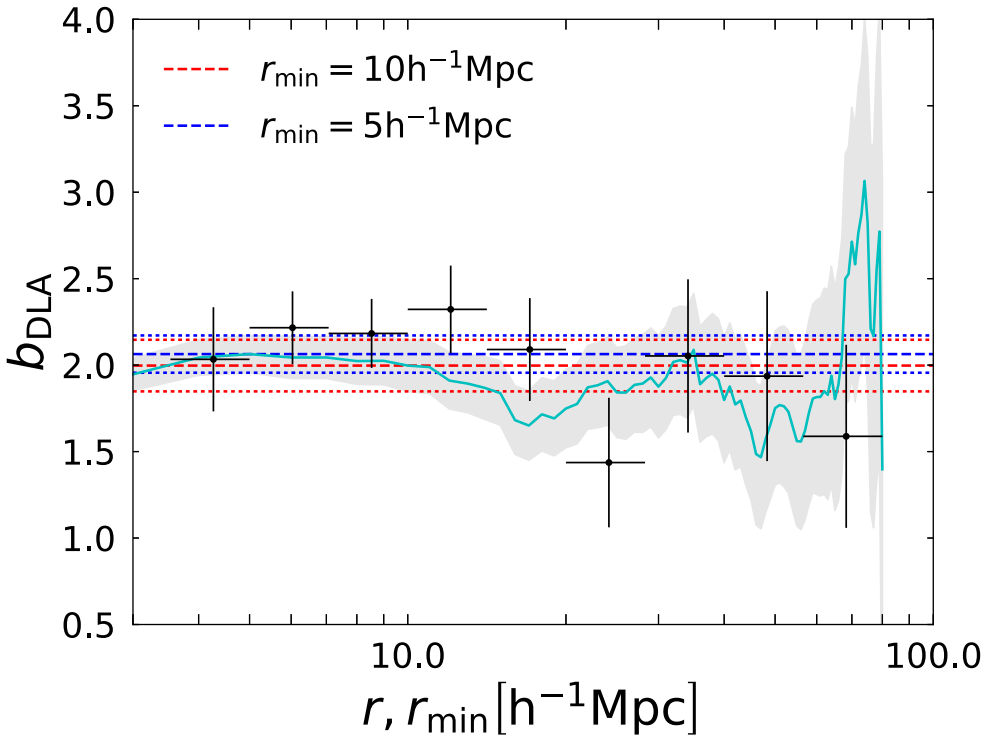

Figure 5. DLA bias versus r obtained by fitting the sample A cross-correlation in the bins in $r = \sqrt{r_\parallel^2 + r_\perp^2}$ indicated by the horizontal errorbars. Dashed lines show values obtained by fitting the whole radial range with two different $r_{\rm min}$, with the $1-\sigma$ error indicated by the dotted lines. The cyan solid line shows $b_{\rm DLA}$ as a function of $r_{\rm min}$, the minimum radius of bins included in the fit. The maximum value of r is fixed at $90\,h^{-1}$ Mpc. Grey area shows the 1σ confidence levels around the cyan line.

Table 4. Summary of all the effects contributing to the difference from the result of FR12 and the final result obtained here for $b_{\rm DLA}$. Errors are obtained from the covariance matrix. The intermediate result after applying the distortion matrix correction needs to be compared to the fiducial result of FR12, $b_{\rm DLA} = 2.17 \pm 0.20$, to see that our methods produce very similar results when applied to the same data with the same fitting model. All our results are for the C2 sample.

Introduced correction	$b_{\rm DLA}$	χ^2(d.o.f)
Original FR12, NOCOR	2.00 ± 0.19	
Use BAOFIT for model fitting	2.01 ± 0.17	
Use our ξ_A, C_{AB}	1.94 ± 0.15	
Distortion matrix correction	2.14 ± 0.16	
(Original FR12, fiducial)	2.17 ± 0.20	
From DR9 to DR12	2.02 ± 0.09	
Change to Planck-2016 cosmological model	1.80 ± 0.08	
Bautista et al. (2017) Ly α bias factors	2.05 ± 0.09	2954.41 (2,864-1)
Smoothing correction G	2.06 ± 0.09	2952.14 (2,864-1)
Smoothing corrections $G \cdot S$	2.08 ± 0.09	2947.73 (2,864-1)
HCD correction with final Ly α bias factors	1.99 ± 0.09	2950.26 (2,864-1)
HCD and metal corrections	2.01 ± 0.09	2954.12 (2,864-1)

We conclude that the difference due to the estimator and covariance matrix (comparing again to the NOCOR case of FR12) is that the DLA bias we obtain is ∼0.07 lower, or reduced by 3.5 per cent, and the error is 20 per cent smaller, compared to FR12. This must be caused by the different way of fitting the continuum to obtain the Ly α transmission, the correction of detected DLAs in the data, and the different covariance matrix we use. We note that of the 20 per cent reduction in the error, 10 per cent is due to the different fitting code as found above.

6.3 Correction for the continuum fitting distortion

Next, we include the correction for the continuum fitting distortion. In FR12, the inclusion of their MTC correction to the fitted model modified the derived bias from $b_{\rm DLA} = 2.00 \pm 0.19$ for their NOCOR case, to $b_{\rm DLA} = 2.17 \pm 0.20$, which was the fiducial or main result in that paper. In our case, using the BAOFIT code and our own estimate of the cross-correlation and covariance matrix, including the distortion matrix method introduced by Bautista et al. (2017,

see Section B for a more detailed explanation) raises the derived bias from $b_{\rm DLA} = 1.94 \pm 0.15$ to $b_{\rm DLA} = 2.14 \pm 0.16$.

We therefore conclude that the two different corrections for the distortions introduced by continuum fitting are very similar. The difference in the derived bias factor when we combine the effects of the fitting method, the estimation of the cross-correlation and covariance matrix, and the continuum fitting distortion corrections, is reduced to only 0.03, and our error based on the covariance matrix is 20 per cent smaller than in FR12 for the reasons discussed in the previous subsections. The fact that two completely independent methods to correct continuum fitting distortions are in good agreement increases our confidence in the accuracy of this correction.

6.4 Larger DR12 data set versus DR9

We now change the data set from DR9 to DR12, using as before sample C2 with $r_{\rm min} = 5\,h^{-1}$ Mpc, and our method to evaluate the cross-correlation and covariance matrix, and the distortion matrix to correct for the continuum fitting effect. The result is that the DLA bias decreases from $b_{\rm DLA} = 2.14 \pm 0.16$ for DR9, to $b_{\rm DLA} = 2.02 \pm 0.09$ for DR12. This change between the two data samples is consistent with the expected statistical error, and the decrease in the errorbar is as expected from the increase of the sample size.

The increased size of the sample from DR9 to DR12 has therefore caused a decrease of the measured DLA bias of 0.75 times the error we infer for DR9. However, we shall now see that systematic differences in the model of FR12 and our own cause larger changes on $b_{\rm DLA}$.

6.5 Cosmological model

Next, we repeat the fit to $b_{\rm DLA}$ for the C2 sample of DR12, applying as before our distortion matrix, and we change the cosmological model from the one used in FR12 based on WMAP results (with parameters $\Omega_m = 0.281$ and $\sigma_8 = 0.8$) to the Planck model we use here, with $\Omega_m = 0.3156$ and $\sigma_8 = 0.831$. The bias changes from $b_{\rm DLA} = 2.02 \pm 0.09$ to $b_{\rm DLA} = 1.80 \pm 0.08$.

There are two main reasons for this change. First, the normalization of the power spectrum at the reference redshift $z_{\rm ref} = 2.3$, which is close to the mean redshift where the DLA-Ly α cross-correlation is measured, is proportional to $b_{\rm DLA}$ times the square of the rms density fluctuation at $z_{\rm ref}$. This fluctuation is usually expressed in terms of its average over a sphere of radius $8\,h^{-1}$ Mpc, which we find to be $\sigma_8(z_{\rm ref}) = 0.3120$ for the FR12 model, and $\sigma_8(z_{\rm ref}) = 0.3161$ in our Planck model. This therefore implies a reduction of $b_{\rm DLA}$ by a factor $(0.3120/0.3161)^2 = 0.974$, if we neglect the small change in $\beta_{\rm DLA}$ corresponding to a change in $b_{\rm DLA}$. We note that the scale of a sphere of $8\,h^{-1}$ Mpc radius is close to the effective scale at which our cross-correlation is measured, so apart from the normalization parameter $\sigma_8(z_{\rm ref})$, there is little variation of $b_{\rm DLA}$ due to the small change in the shape of the power spectrum between the two models. For instance, a 2σ change in n_s does not significantly change the value of the recovered bias.

The second reason is the change in the angular diameter distance and Hubble constant. Our measurements of the cross-correlation are made at known angular and redshift separations, whereas the model correlation function is predicted in comoving coordinates in units of h^{-1} Mpc. The ratio of the quantity $H_0 D_A(z_{\rm ref})$ in the model used in this paper and the FR12 model is 0.9690, and the ratio of the quantity $H_0/H(z_{\rm ref})$ for our model and the FR12 model is 0.9484. We take an average of these two scaling factors, ∼0.96, as the characteristic ratio by which the comoving scale that is computed from observed angular and redshift separations changes between the two cosmological models. The model ΛCDM cross-correlation varies approximately as $\xi \sim r^{-2}$ over the range of scales in which our measurement is most significant, so this implies an approximate reduction in the inferred $b_{\rm DLA}$ by a factor of ∼0.92. Combining this with the previous reduction factor from $\sigma_8(z_{\rm ref})$, we see how a total reduction of the inferred $b_{\rm DLA}$ by ∼10 per cent due to the change of the cosmological model is explained.

6.6 Ly α forest bias parameters

Apart from the cosmological model, our result on $b_{\rm DLA}$ is also strongly affected by the Ly α forest bias parameters. The bias parameters used in FR12 were $\beta_{{\rm Ly}\alpha} = 1$ and $b_{{\rm Ly}\alpha}(1 + \beta_{{\rm Ly}\alpha}) = -0.336$ at a reference redshift $z_{\rm ref} = 2.25$, taken from Slosar et al. (2011). This needs to be transformed to the reference redshift we use of $z_{\rm ref} = 2.3$, using the assumed evolution of the Ly α forest bias of $b_{{\rm Ly}\alpha} \propto (1+z)^{2.9}$ in all the papers that have measured the Ly α autocorrelation. The result is $b_{{\rm Ly}\alpha}(1 + \beta_{{\rm Ly}\alpha}) = -0.351$.

These values were updated first by the analysis of Blomqvist et al. (2015), who fitted the DR11 Ly α autocorrelation applying the linear theory model only to scales $r > 40\,h^{-1}$ Mpc. Their result was $b_{{\rm Ly}\alpha}(1 + \beta_{{\rm Ly}\alpha}) = -0.374$, and $\beta_{{\rm Ly}\alpha} = 1.39$, at $z_{\rm ref} = 2.3$. Using these Ly α bias parameters, the DLA bias changes only from $b_{\rm DLA} = 1.80 \pm 0.08$ to $b_{\rm DLA} = 1.82 \pm 0.08$. However, the more recent analysis of DR12 by Bautista et al. (2017) gives a substantially different result, when fitting all the data down to $r > 10\,h^{-1}$ Mpc to the same model as before (model labelled Ly α in their table 5): $b_{{\rm Ly}\alpha}(1 + \beta_{{\rm Ly}\alpha}) = -0.326$, and $\beta_{{\rm Ly}\alpha} = 1.246$, which makes our result for the DLA bias increase to $b_{\rm DLA} = 2.05 \pm 0.09$. The reason for this change from Blomqvist et al. (2015) is that the Ly α autocorrelation data prefer a lower Ly α bias factor at smaller scales, and in fact Bautista et al. (2017) noted that this simple model does not provide a good fit to the whole radial range.

The dependence of our result on the Ly α forest bias parameters can be understood by noting that the amplitude of the cross-correlation model is proportional to $\sigma_8^2(z_{\rm ref})\, b_{{\rm Ly}\alpha}\, b_{\rm DLA}$. Only this product can be inferred from the cross-correlation measurement. However, the angular dependence of the redshift distortion factors introduces a more complex dependence on $\beta_{{\rm Ly}\alpha}$ and $\beta_{\rm DLA}$. We show in Fig. 6 the inferred $b_{\rm DLA}$ as a function of $\beta_{{\rm Ly}\alpha}$, when keeping $b_{{\rm Ly}\alpha}(1 + \beta_{{\rm Ly}\alpha})$ fixed, as represented by the red circles (black circles include the HCD correction and are discussed below). Errorbars are our statistical errors from the cross-correlation measurement. Following FR12, we fit a power-law dependence $b_{\rm DLA} \propto \beta_{{\rm Ly}\alpha}^{\gamma}$, finding $\gamma = 0.21$ over the range of interest shown in Fig. 6.

The variation of $b_{\rm DLA}$ is nearly proportional to $b_{{\rm Ly}\alpha}^{-1}$ at fixed $\beta_{{\rm Ly}\alpha}$, except for the fact that $\beta_{\rm DLA} \propto b_{\rm DLA}^{-1}$, implying that $b_{\rm DLA}$ increases a bit faster than expected with decreasing $b_{{\rm Ly}\alpha}$ because of the need to compensate for a smaller redshift distortion factor.

6.7 New additions to the cross-correlation model

Our model incorporates improvements that were not present in FR12: the correction for binning of the cross-correlation and the wavelength PSF, and the HCD correction. The size of the bins in $r_\parallel$ and $r_\perp$ of $2\,h^{-1}$ Mpc is corrected for by multiplying by the function $G(\boldsymbol{k})$ in Fourier space (see Section 4 and equation 11). This increases our last value $b_{\rm DLA} = 2.05 \pm 0.09$ for the linear Ly α model of Bautista et al. (2017) to $b_{\rm DLA} = 2.06 \pm 0.09$. The

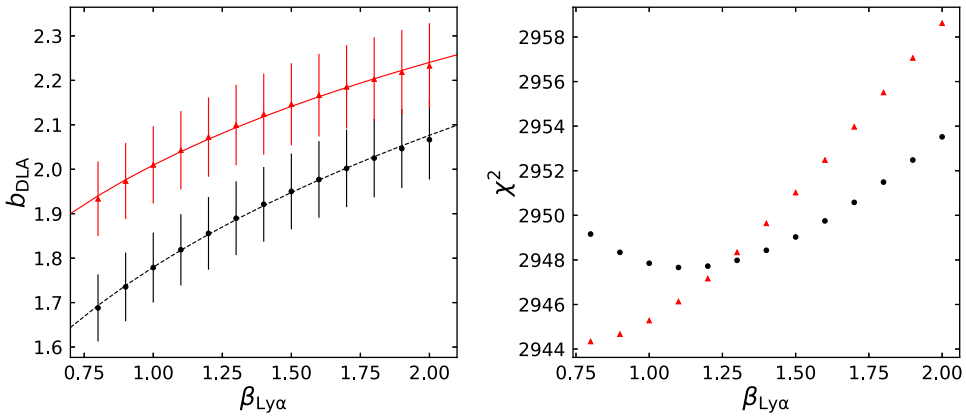

Figure 6. Left: Inferred $b_{\rm DLA}$ as a function of $\beta_{{\rm Ly}\,\alpha}$ when keeping the fixed value $b_{{\rm Ly}\,\alpha}(1+\beta_{{\rm Ly}\,\alpha})=-0.325$. Red triangles do not include the HCD correction, and black circles include it. Also shown are power-law fits $b_{\rm DLA}\propto\beta_{{\rm Ly}\,\alpha}^{\gamma_1}$, with $\gamma_1=0.21$ for no HCD correction (solid red line) and $\gamma_1=0.23$ with the correction (dashed black line). Right: Values of χ^2 when fitting $b_{\rm DLA}$ for different $\beta_{{\rm Ly}\,\alpha}$, with the HCD correction (black circles) and without (red triangles).

wavelength PSF includes the BOSS spectrograph resolution and the rebinning of the BOSS spectral pixels into analysis pixels that are three times wider, as discussed in Section 4, and is corrected by multiplying by the function $S(k_\parallel)$. This further increases our result to $b_{\rm DLA}=2.08\pm0.09$.

A more important impact on the DLA bias is caused by the HCD correction, introduced by Bautista et al. (2017) and discussed in Section 4. The black crosses in Fig. 6 show the inferred $b_{\rm DLA}$ as a function of β_α when the HCD correction is included, and the power-law fit for this case, with $\gamma=0.23$, is shown as the dashed black line. At a fixed value of β_α, including the HCD correction causes a reduction of ∼10 per cent on $b_{\rm DLA}$. However, this correction must be included self-consistently with the parameters fitted to the Ly α autocorrelation. We therefore change to the final values of the Ly α forest bias factors we use for our fiducial result of the DLA bias in this paper, those in table 3 of Bautista et al. (2017): $b_{{\rm Ly}\alpha}(1+\beta_{{\rm Ly}\alpha})=-0.326$ and $\beta_{{\rm Ly}\alpha}=1.663$, which were obtained by including not only the HCD correction in the fitted model but also an additional correction due to metal lines. For these values, we find $b_{\rm DLA}=2.18\pm0.10$ when not including the HCD correction in the cross-correlation fit, and $b_{\rm DLA}=1.99\pm0.09$ when including it (corresponding to the red and black curves in Fig. 6, respectively, at $\beta_{{\rm Ly}\alpha}=1.663$). We have generally not included the correction for metal lines in this paper, because their effect is not detected in our DLA-Ly α cross-correlation. However, we find that including the same metal-line correction in our analysis increases the DLA bias to $b_{\rm DLA}=2.01\pm0.09$, and worsens the χ^2 value.

The dependence of $b_{\rm DLA}$ on the HCD correction is therefore substantially smaller than 10 per cent when we use self-consistently the values of the Ly α bias factors that fit the Ly α autocorrelation. The reason why $\beta_{{\rm Ly}\alpha}$ needs to increase when including the HCD correction is that the latter adds to the cross-correlation model a function that is elongated in the parallel direction, accounting for the Voigt profiles with damped wings of HCD absorbers. This needs to be compensated by an increased Kaiser effect in the linear model, causing a tangential elongation. The change in $b_{\rm DLA}$ from the model with $\beta_\alpha=1.246$ and no HCD correction, to the model with $\beta_\alpha=1.663$ with the HCD correction, is less than 5 per cent (from 2.08 to 1.99), and reflects the true impact of the HCD correction.

Finally, the right-hand panel of Fig. 6 shows the χ^2 value of our fit as a function of $\beta_{{\rm Ly}\alpha}$, with no HCD correction (red triangles) and including it (black circles). It is interesting that the best-fitting value for no HCD correction, $\beta_{{\rm Ly}\alpha}=1.1\pm0.3$, is lower than that obtained by Bautista et al. (2017) (although only at the $1.5-\sigma$ level with the HCD correction), and that the HCD correction worsens our fit by $\Delta\chi^2\simeq4$. This is probably an indication that the HCD correction is not a sufficiently good model of the impact of HCD absorption wings in the Ly α spectra.

To summarize all the differences from FR12 and model dependences discussed in this section, Table 4 lists all the changes of $b_{\rm DLA}$ caused by each of the effects we have discussed.

7 DISCUSSION

7.1 Systematic errors: cross-correlation modelling

So far, all of the errors we have quoted for $b_{\rm DLA}$ include only statistical errors of the cross-correlation measurement, computed either from our covariance matrix or the bootstrap analysis. We now discuss systematic errors, which arise from two sources: uncertainties in the model of the cross-correlation to be used in the fit, and impurity of the DLA sample. We discuss first the modelling uncertainties.

There are several possible sources of systematic error of $b_{\rm DLA}$ in our modelling procedure: the continuum fitting correction, the assumed cosmological model, the use of linear theory, the Ly α bias factors and the HCD correction. We believe our continuum fitting correction is accurate, in view of the good agreement of two independent methods of applying this correction from FR12 and the distortion matrix procedure used here (see Section 6.3), and the tests that have been made with mocks (Bautista et al. 2017). While we have shown that there is a high sensitivity to the cosmological model (a 10 per cent variation of $b_{\rm DLA}$ is caused by the update from the WMAP model of FR12 to the Planck model we use), this does not cause a large systematic if we believe that the results of Planck Collaboration (2016) are accurate. As we shall see below, we are particularly interested in systematics that might lower our inferred value of $b_{\rm DLA}$, to bring it in closer agreement with expectations from cosmological simulations of galaxy formation, and this can only occur by further increasing $\sigma_8(z_{\rm ref})$ or Ω_m in the cosmological model. Linear theory seems well justified from the constant value of $b_{\rm DLA}$ with scale (Fig. 5) and the large value of $r_{\rm min}$ we are using, although precise predictions of non-linearities in the DLA-Ly α cross-correlation from cosmological simulations would be highly desirable to test this.

We believe the more important sources of systematics are in the uncertainties in the Ly α bias factors and the HCD correction determined from the Ly α autocorrelation. If we use the two models fitted to the Ly α autocorrelation in Bautista et al. (2017), the 'Lyα' one in their table 5 without HCD correction and $\beta_{\mathrm{Ly}\,\alpha} = 1.246$, and the full model in their table 3 with HCD correction and $\beta_{\mathrm{Ly}\,\alpha} = 1.663$, the difference of $\sim$5 per cent in the implied b_{DLA} between the two models is a good estimate of our systematic error, which is comparable to our statistical error of b_{DLA}. The model including the HCD correction gives the lowest value of b_{DLA}, and we conservatively take it as our final result to compare with predictions from galaxy formation simulations. The statistical errors in the Ly α bias factors of the models of Bautista et al. (2017) are negligibly small for our purpose.

We note that even though the HCD correction is not a very accurate representation of the true effect of HCDs, since it does not take into account the precise Voigt profile shape of the absorbers, we believe the important thing is that the same model that provides a good fit to the measured Ly α autocorrelation is used to model the DLA-Ly α cross-correlation.

7.2 Systematic errors: sample purity and selection

We now discuss the purity and selection effects of our samples A, C1 and C2. DLAs are detected using the automatic algorithm described in Noterdaeme et al. (2009, 2012), which searches for regions of strong absorption that are consistent with a DLA line plus the random absorption by the Ly α forest. We therefore expect that some fraction of these candidate DLAs in the catalogue are not real DLAs, and are likely to be instead regions of strong absorption over a sufficiently broad velocity interval to look approximately like a damped profile, but with a real H I column density much less than 10^{20} cm^{-2}. These false detections should increase at low $N_{\mathrm{H\,I}}$ and low signal-to-noise ratio. For very low signal-to-noise ratio, some false DLAs may not correspond to any absorber but be mostly caused by noise, but these cases should be rare with our imposed cut of CNR > 3.

The fact that we observe no variation of b_{DLA} with column density suggests that the effect of these contaminants on b_{DLA} is not large. Either the impurity is too small to affect our result, or the false DLAs are regions of absorption with a bias that is close to that of DLAs. An additional argument against a large level of impurity of our sample is the result of Mas-Ribas et al. (2017) on the dependence of the mean equivalent width of high-ionization lines of DLAs on $N_{\mathrm{H\,I}}$ (see their fig. 11), which varies only by 20 per cent over the available range of column densities. Moreover, this small variation is not necessarily due to a variation of the impurity level, but can be caused by a real physical effect.

A more detailed study requires predicting the purity of our catalogue using Ly α forest mock spectra. This is not a simple calculation, because the mock spectra must have the correct distribution of HCDs with broad absorption features that can mimic DLAs (which our current mocks are not designed to reproduce), and DLAs must be inserted in the mock spectra with the correct cross-correlation, so we leave this for future studies.

Different issues arise with DLAs detected bluewards of the Ly β quasar emission line, where the Ly β forest is superposed with Ly α absorption and the possibilities of confusion increase. Often, a DLA may be detected in part because there is a Ly β line from an absorber that has most of the column density, resulting in an incorrect assigned redshift. In the absence of any real Ly α absorber, these absorbers with incorrect redshifts should contribute a zero cross-correlation to our measurement, decreasing the fitted value of b_{DLA} by a fractional amount equal to the fraction of these systems. Our sixth cut, applied to define sample A, eliminates these misidentified objects, which are therefore present only in our samples C1 and C2 (see Section 2; sample C1 is the largest, due to not including cuts 4 and 5 that eliminate DLAs too close to the O VI and Ly α quasar emission lines).

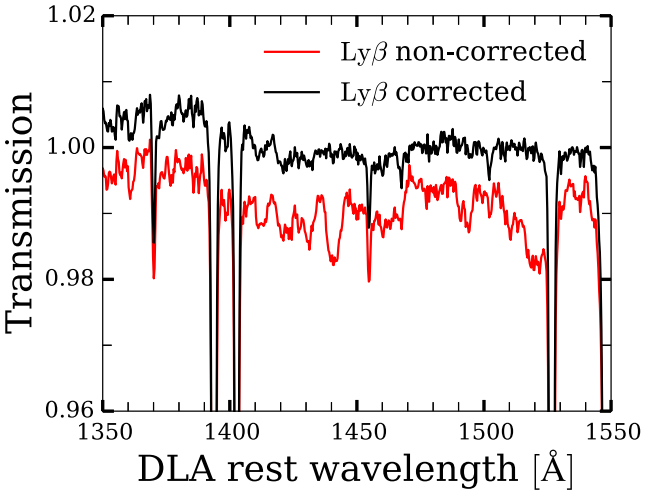

Figure 7. DLA transmission spectrum of the total sample of Mas-Ribas et al. (2017) (black line), and the same transmission spectrum for the larger sample including DLAs in the Ly β forest region, which may include Ly β absorbtion lines misidentified as Ly α lines of DLAs (thick, red line). The absorption feature at $\sim$1441 Å suggests that these misidentified DLAs are $\sim$1 per cent of the sample.

For $r_{\mathrm{min}} = 5\,h^{-1}$ Mpc, we measure $b_{\mathrm{DLA}} = 2.06 \pm 0.11$ for sample A, $b_{\mathrm{DLA}} = 1.97 \pm 0.08$ for sample $C1$ and $b_d = 1.99 \pm 0.09$ for sample C2. These errors are obtained from our covariance matrix and are roughly proportional to the inverse square root of the number of DLAs in each sample. The variation in the bias of the DLA samples is hardly significant, especially if the larger bootstrap errors are considered (see Table 2), although they go in the expected direction of a decreased b_{DLA} in samples with an expected higher impurity. This suggests that DLAs removed by cuts four to six are primarily DLAs or other absorption systems with a similar bias factor.

The level of impurity due to Ly β absorption lines confused by Ly α n samples C1 and C2 can also be estimated by stacking the absorption spectra. If a Ly β absorption line is incorrectly attributed to Ly α absorption at $\lambda_\alpha = 1216$ Å, then the true Ly α absorption will appear at $32/27\lambda\alpha = 1441$ Å. Fig. 7 shows the stacked spectrum obtained with the technique of Mas-Ribas et al. (2017) for what was designated as 'total sample' by these authors, which excluded DLAs found in the Ly β forest region, as the black line. The red line is the stacked spectrum of the larger sample including DLAs in the Ly β forest region, and shows the expected absorption of misidentified absorbers at a level of $\sim$ 1 per cent. We take this as an upper limit to the fraction of systems in our C1 and C2 samples that are Ly β lines wrongly identified as Ly α lines of DLAs, because these Ly β lines are more likely to be identified when they have superposed true Ly α absorption. This indicates that this contamination of the sample is very small and not significant compared to our statistical errors.

In general, the inclusion of any false absorbers in our catalogue arising purely from noise or from misidentified redshifts can only decrease our measured b_{DLA}, because the false absorbers have a null

Table 5. Number of DLAs in the subsamples ZC1 to ZC3 and NC1 to NC3, drawn from sample C1.

Name	Number of DLAs
ZC1	6319
ZC2	6664
ZC3	10 359
NC1	8 613
NC2	7 788
NC3	6 941

average contribution to the cross-correlation. The presence of HCD absorbers misidentified as DLAs may increase our measured bias only if the HCD bias is higher than that of DLAs. There is, however, a systematic arising from a selection effect that may increase the measured $b_{\rm DLA}$, already mentioned in FR12: if DLAs are more likely to be detected when the Ly α forest that is superposed with their damped wings is weaker than average, then this would preferentially select DLAs surrounded by high-density large-scale regions, over those in low-density regions. The reason is that the DLA-Ly α cross-correlation is negative along the line of sight at $|r_\parallel| \gtrsim 20\,h^{-1}$ Mpc owing to redshift space distortions, implying weaker Ly α forest absorption over the damped wings of DLAs in more overdense regions. This is a selection effect that can only be properly corrected with the use of adequate mock spectra with DLAs inserted with the correct cross-correlation with the Ly α forest. Again, we believe this correction is unlikely to be large because of the absence of dependence of $b_{\rm DLA}$ on $N_{\rm H\,I}$, but future studies will need to better address this question.

7.3 Evolution of the bias factor

The lack of a significant dependence of $b_{\rm DLA}$ on redshift and column density was shown for sample A in Fig. 4, although a large scatter was noticed for the redshift dependence for the case with $r_{\rm min} = 10\,h^{-1}$ Mpc, with a lower bias for the middle redshift than the low and high redshift ones by $\sim 2.5\sigma$. To explore if this scatter might indicate something other than noise, we repeat the measurement using sample C1, taking into account that decreased purity is unlikely to be very important as argued in Section 7.2.

We define six new subsamples by dividing the C1 sample into the same three redshift and column density bins as in Table 1 for sample A. The number of systems in the new subsamples are shown in Table 5, nearly doubling those from sample A. Results are shown in Fig. 4, where squares with dashed errorbars show the bias values obtained with the new subsamples, and the circles with solid errorbars show the previous results from sample A. The normal scatter for sample C1 suggests that the anomalously high scatter in sample A is only due to statistical noise.

The fact that no change of the bias factor (within 10 per cent) is seen between redshift 2–3 suggests that the characteristic host halo mass is decreasing with redshift. The independence with column density also suggests that the mean NH I radial profile is similar in host haloes of different masses.

7.4 Implications on the distribution of DLA host halo masses

The bias factor of dark matter haloes as a function of their mass is robustly predicted in analytic models and numerical simulations (see e.g. Sheth & Tormen 1999; Tinker et al. 2010), and therefore our derived DLA bias factor implies a condition on the characteristic mass of haloes hosting DLAs. We use the model of Tinker et al. (2010) to calculate the halo bias at the mean redshift of our cross-correlation measurement $z = 2.3$, shown as the thick solid curve in Fig. 8 (both left-hand and right-hand panels). The grey horizontal line with the shaded band is the value of $b_{\rm DLA}$ for our C1 sample and $r_{\rm min} = 5\,h^{-1}$ Mpc, with the bootstrap error. This is the result with the smallest error that we believe we can trust, as we have argued above. However, it does not include the systematic error arising mainly from the Ly α bias factors and impurities in the catalogue.

If all DLAs were in haloes with a single value of the mass, the inferred halo mass would be placed from $2.5 \times 10^{11}\,h^{-1}\,{\rm M}_\odot$ to $5 \times 10^{11}\,h^{-1}\,{\rm M}_\odot$ given our statistical errorbar, corresponding to a massive galaxy. However, in any realistic model, DLAs host haloes should have a broad mass range. Our measurement yields only the mean bias factor, which depends on the DLA cross-section as a function of halo mass as discussed in FR12. This cross-section depends on the distribution of gas in haloes, and therefore, on the complex physics of gas accretion, galaxy formation, and galactic and quasar winds that can expel gas from a central galaxy to the outer regions of haloes or to the intergalactic medium.

Following FR12, we assume a power-law distribution of the DLA cross-section $\Sigma(M_h)$ as a function of halo mass,

$$\Sigma(M_h) = \Sigma_0 \left(\frac{M_h}{M_{\rm min}}\right)^{\alpha} \quad (M > M_{\rm min})\,. \tag{15}$$

The predicted mean DLA bias under this simple assumption is

$$b_{\rm DLA} = \frac{\int_{M_{\rm min}}^{M_{\rm max}} n(M_h)\Sigma(M_h)b(M_h)\,{\rm d}M_h}{\int_{M_{\rm min}}^{M_{\rm max}} n(M_h)\Sigma(M_h)\,{\rm d}M_h}, \tag{16}$$

where $n(M_h)$ is the number density and $b(M_h)$ the bias of haloes of mass M_h. We have also assumed that the cross-section is negligible below a minimum mass $M_{\rm min}$ and above a maximum mass $M_{\rm max}$. Numerical simulations of galaxy formation including hydrodynamics and complex recipes for star formation and galaxy winds driven by supernova explosions have been extensively studied by several groups (e.g. Pontzen et al. 2008; Tescari et al. 2009; Bird et al. 2014), and can predict this relationship. In particular, Bird et al. (2014) find a relation that is well fitted by a power law over the halo mass range that can be probed by their simulations, $10^{8.5}\,h^{-1}\,{\rm M}_\odot < M < 10^{12}\,h^{-1}\,{\rm M}_\odot$. At lower masses, the intergalactic photoionized gas has sufficient pressure to slow the accretion on to haloes. At higher halo masses, the small box of their simulations do not allow enough haloes to be included to derive solid predictions.

The mean DLA bias computed in this model is shown in the left-hand panel of Fig. 8 as a function of $M_{\rm min}$, and in the limit of infinite $M_{\rm max}$, for six different values of the power-law index α. For $M_{\rm min} = 10^{8.5}\,h^{-1}\,{\rm M}_\odot$, the required slope to match our DLA bias is $\alpha \sim 1.1$. The simulations of Bird et al. (2014) tend to give a lower slope of $\alpha \simeq 1.0$, which corresponds to $b_{\rm DLA} \simeq 1.7$ for the same value of $M_{\rm min}$, which is discrepant with our measurement at the $\sim 2-\sigma$ level if we use only our statistical, bootstrap errors, but may be more consistent with what we measure when including plausible systematic errors from uncertainties in the Ly α forest bias factors and the effects of catalogue impurity. If a bias factor $b_{\rm DLA} > 1.9$ is confirmed by future by improved determinations, then either a steeper slope or higher value of $M_{\rm min}$ compared to the Bird et al. (2014) simulations would be required. Results of earlier simulations that had weaker galactic winds and predicted slopes of $\alpha \simeq 0.7$ (Pontzen et al. 2008), with much lower implied DLA bias

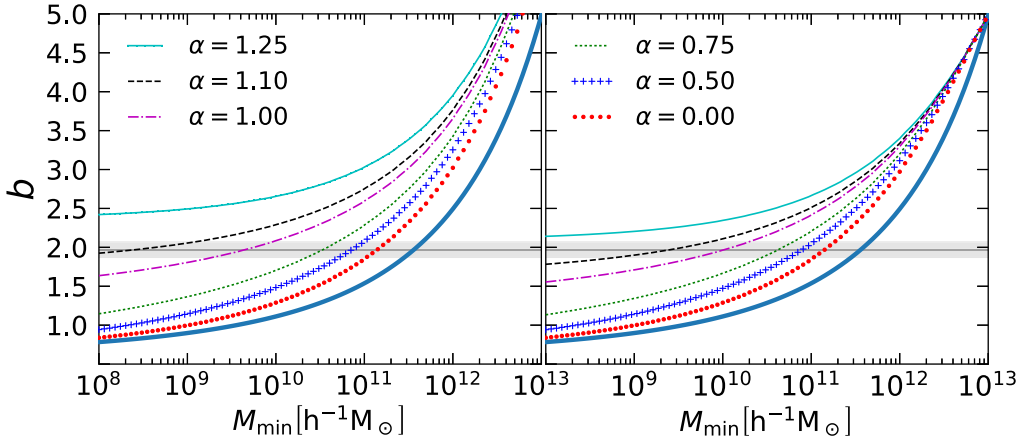

Figure 8. Left: Average DLA bias when the DLA cross-section as a function of halo mass follows the power-law relation in equation (15), for the indicated values of the power-law index α, as a function of the lower mass cut-off $M_{\rm min}$. Right: Same as in the left-hand panel but including an upper mass cut-off at $M_{\rm max} = 10^{13}\,h^{-1}\,{\rm M}_{\odot}$. The bias for a single halo mass is shown in both panels as the thick solid line. All cases are computed at $z = 2.3$. Horizontal shaded region is our derived value for the DLA bias (sample C1, $r_{\rm min} = 5\,h^{-1}$ Mpc) and the 1σ statistical bootstrap error (not including errors on the Lyα forest bias factors).

factors of $b_{\rm DLA} < 1.5$, are strongly ruled out by our measurement. In general, the high value of $b_{\rm DLA}$ we measure implies that the structure of DLAs is affected by strong galactic winds, which are able to decrease the cross-sections in low-mass haloes by expelling gas to the intergalactic medium, and increase cross-sections in high-mass haloes by spreading gas out to large radius.

However, Bird et al. (2014) assumed an extrapolation of their power-law fit to $M_h > 10^{12}\,h^{-1}\,{\rm M}_{\odot}$ to derive a bias factor $b_{\rm DLA} \simeq 1.7$ from their fitted power-law slope, because their small simulations cannot predict the properties of the rare, more massive haloes. These massive haloes are very highly biased and make an important contribution to the mean bias of DLAs. The right-hand panel of Fig. 8 shows the same models using now an upper cut-off $M_{\rm max} = 10^{13}\,h^{-1}\,{\rm M}_{\odot}$, and we see that in this case, the predicted bias factor for $M_{\rm min} = 10^{8.5}\,h^{-1}\,{\rm M}_{\odot}$ and $\alpha = 1$ already decreases to $b_{\rm DLA} \simeq 1.6$. The results of Bird et al. (2014) can therefore agree with our measurement at better than $2-\sigma$ only if we allow for a systematic error and if there is no substantial flattening of the slope of the $\Sigma(M_h)$ relation at $M_h > 10^{12}\,h^{-1}\,{\rm M}_{\odot}$. There are reasons to expect this flattening of the slope because at low redshift, we observe that massive haloes are associated with galaxy groups and clusters containing most of the baryons in X-ray emitting hot gas, where much of the cold gas in galaxies is destroyed owing to tidal and ram-pressure stripping (Fabian 2012). However, at the redshifts where DLAs in BOSS are found, the amount of cold gas in very massive haloes is not well known. Larger and better simulations, and observations of galaxy clusters at high redshift, are required to clarify this question.

Our improved measurement of the bias factor of DLAs has an impact on forecasts for 21-cm surveys of H I galaxies: a higher value of the bias implies a larger amplitude of the 21-cm fluctuations (see e.g. Chang et al. 2010; Chang & GBT-HIM Team 2014; Castorina & Villaescusa-Navarro 2017; Villaescusa-Navarro et al. 2016). Our new value is very similar to the previous one by FR12, with a reduced error and a more detailed analysis of model dependences. The 21-cm fluctuation amplitude depends on the mean bias of all absorbers weighted by their H I column density. The lack of dependence of $b_{\rm DLA}$ on $N_{\rm H\,I}$, and the fact that most of the known neutral hydrogen in the Universe resides in DLAs, strongly suggest that our derived value $b_{\rm DLA} \simeq 2.0$ should apply for the neutral gas that will be detected in 21-cm surveys. Although these surveys should include dust-absorbed systems that are not included in our DLA sample and lower column density systems that we also do not include, it is difficult that these systems may change the mean bias factor appreciably.

Finally, we comment on one theoretical aspect of the bias of dark matter haloes that may influence the comparison of the theoretically predicted and observed DLA bias factor. The halo bias is not only a function of mass, but also of the assembly history of a halo, a phenomenon known as 'assembly bias' (e.g. Borzyszkowski et al. 2017). Haloes of a fixed mass in high-density regions tend to have accreted their mass recently, whereas in low-density regions the accretion rate is lower. The DLA cross-section may depend also on the accretion history: for a fixed halo mass, a high accretion rate may imply more atomic gas is available at large radius to give rise to a DLA system, and at the same time, a higher bias owing to the assembly bias effect. This effect may be missed by simulation results like those in Bird et al. (2014) when the bias factor is inferred from the DLA host halo mass distribution and the same type of theoretical relation of bias and halo mass we have used here, instead of being directly obtained from the simulation. Alternatively, one may achieve a steeper slope cross-correlation versus halo-mass relation, and then get a higher predicted DLA bias, by changing some parameters of the winds models in simulation. Future studies should therefore also attempt to include the effect of assembly bias or wind model when comparing to the observational result.

8 SUMMARY AND CONCLUSIONS

We have measured the cross-correlations of DLAs and the Lyα forest for several samples of DLAs of the final DR12 of BOSS: 23 342 DLAs with $N_{\rm H\,I} \geq 10^{20}\,{\rm cm}^{-2}$ in the redshift range of $2.0 \geq z_{\rm DLA} \geq 3.5$. We have found that the simple linear theory model for this cross-correlation, with the redshift distortions predicted by Kaiser (1987), is fully consistent with the data, and we have obtained the DLA bias factor required to match the measured cross-correlation amplitude. Our main conclusions are as follows:

(i) We measure $b_{\rm DLA} = (1.99 \pm 0.11)$ for sample C2, extending the fit range down to $r_{\rm min} = 5\,h^{-1}$ Mpc. A more conservative result,

using sample A and $r_{\rm min} = 10\,h^{-1}$ Mpc to avoid possible non-linear effects, yields $b_{\rm DLA} = (2.00 \pm 0.19)$. Both values are similar to the previous result reported by FR12, but the detailed comparison depends on several differences in the analysis and model dependences discussed in Section 6.

(ii) We do not find any dependence of the DLA bias on redshift and $N_{\rm H\,I}$, at the level of ∼10 per cent over the ranges of $2 < z < 3$ and $20 < \log N_{\rm H\,I} < 21.5$. The independence on redshift suggests that the characteristic host halo mass is decreasing with redshift, and the independence with column density suggests that the mean $N_{\rm H\,I}$ radial profile is similar in host haloes of different masses.

(iii) The value of the DLA bias does not significantly change amongst our samples that include or exclude DLAs in the Ly β forest or near the Ly α and O VI quasar emission lines, suggesting that systematics associated with these cuts are small.

(iv) We detect no scale dependence in the DLA bias, which reinforces the agreement of the measured cross-correlation with the linear model we assume, based on the ΛCDM power spectrum with the parameters determined by Planck Collaboration (2016).

(v) The principal systematic errors that need to be addressed to make the measurement of $b_{\rm DLA}$ more robust are the dependence on the Ly α forest bias parameters and the HCD correction, and the effects of impurities and selection effects in the DLA catalogue. The absence of any dependence on column density, and the small variations of the DLA bias with the HCD correction when used consistently in the same models that fit the Ly α autocorrelation results of Bautista et al. (2017) suggests that these systematics are not larger than our statistical errors. The small variation of the high-ionization lines mean equivalent width with $N_{\rm H\,I}$ found in Mas-Ribas et al. (2017) also suggest the same thing.

(vi) Assuming the DLA cross-section versus halo-mass relation $\Sigma(M_h) \propto M_h^{\alpha}$ down to $M_{\rm min} \sim 10^{8.5}\,h^{-1}\,{\rm M}_{\odot}$, we find that $\alpha > 1$ is required to match the observed $b_{\rm DLA}$, a steeper relation than is predicted in most numerical simulations of galaxy formation. Even for the simulations with strong winds analysed by Bird et al. (2014), which predict a steeper relation than previous models, the implied bias is only marginally consistent with our observational determination, and needs to assume an extrapolation of this power-law relation with $\alpha \simeq 1$ up to halo masses much larger than the ones being probed by their simulation results. The effect of assembly bias may increase the theoretical prediction for $b_{\rm DLA}$ and help bringing it into agreement with our observational determination.

ACKNOWLEDGEMENTS

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the US Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, University of Cambridge, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington and Yale University.

IPR and JME were supported by the Spanish MINECO under projects AYA2012-33938 and AYA2015-71091-P and MDM-2014-0369 of ICCUB (Unidad de Excelencia 'María de Maeztu'). AFR is supported by a STFC Rutherford Fellowship, grant reference ST/N003853/1. SB was supported by NASA through Einstein Postdoctoral Fellowship Award Number PF5-160133.

REFERENCES

Barnes L. A., Garel T., Kacprzak G. G., 2014, PASP, 126, 969
Bautista J. E. et al., 2017, A&A, 603, A12
Bird S., Vogelsberger M., Haehnelt M., Sijacki D., Genel S., Torrey P., Springel V., Hernquist L., 2014, MNRAS, 445, 2313
Bird S., Haehnelt M., Neeleman M., Genel S., Vogelsberger M., Hernquist L., 2015, MNRAS, 447, 1834
Blomqvist M. et al., 2015, JCAP, 11, 034
Bolton A. S. et al., 2012, AJ, 144, 144
Borzyszkowski M., Porciani C., Romano-Díaz E., Garaldi E., 2017, MNRAS, 469, 594
Bovy J. et al., 2011, ApJ, 729, 141
Busca N. G. et al., 2013, A&A, 552, A96
Castorina E., Villaescusa-Navarro F., 2017, MNRAS, 471, 1788
Cen R., 2012, ApJ, 748, 121
GBT-HIM TeamChang T.-C., 2014, Exascale Radio Astronomy, Vol. 2
Chang T.-C., Pen U.-L., Bandura K., Peterson J. B., 2010, Nature, 466, 463
Cole S., Kaiser N., 1989, MNRAS, 237, 1127
Cooke J., Wolfe A. M., Gawiser E., Prochaska J. X., 2006, ApJ, 652, 994
Crighton N. H. M. et al., 2015, MNRAS, 452, 217
Dawson K. S. et al., 2013, AJ, 145, 10
Delubac T. et al., 2015, A&A, 574, A59
Eisenstein D. J. et al., 2011, AJ, 142, 72
Fabian A. C., 2012, ARA&A, 50, 455
Font-Ribera A. et al., 2012, JCAP, 11, 59 (FR12)
Font-Ribera A. et al., 2014, JCAP, 5, 27
Fox A. J., Ledoux C., Petitjean P., Srianand R., 2007a, A&A, 473, 791
Fox A. J., Petitjean P., Ledoux C., Srianand R., 2007b, A&A, 465, 171
Fumagalli M., Prochaska J. X., Kasen D., Dekel A., Ceverino D., Primack J. R., 2011, MNRAS, 418, 1796
Gunn J. E. et al., 1998, AJ, 116, 3040
Gunn J. E. et al., 2006, AJ, 131, 2332
Haehnelt M. G., Steinmetz M., Rauch M., 1998, ApJ, 495, 647
Jorgenson R. A., Murphy M. T., Thompson R., 2013, MNRAS, 435, 482
Kaiser N., 1987, MNRAS, 227, 1
Kirkby D. et al., 2013, JCAP, 3, 24
Kirkpatrick J. A., Schlegel D. J., Ross N. P., Myers A. D., Hennawi J. F., Sheldon E. S., Schneider D. P., Weaver B. A., 2011, ApJ, 743, 125
Kulkarni V. P., Fall S. M., 2002, ApJ, 580, 732
Kulkarni V. P., Fall S. M., Lauroesch J. T., York D. G., Welty D. E., Khare P., Truran J. W., 2005, ApJ, 618, 68
Lochhaas C. et al., 2016, MNRAS, 461, 4353
McDonald P., Miralda-Escudé J., 1999, ApJ, 519, 486
McDonald P. et al., 2006, ApJS, 163, 80
Mas-Ribas L. et al., 2017, ApJ, 846, 4
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Møller P., Fynbo J. P. U., Ledoux C., Nilsson K. K., 2013, MNRAS, 430, 2680
Neeleman M., Wolfe A. M., Prochaska J. X., Rafelski M., 2013, ApJ, 769, 54
Neeleman M., Prochaska J. X., Wolfe A. M., 2015, ApJ, 800, 7
Noterdaeme P., Petitjean P., Ledoux C., Srianand R., 2009, A&A, 505, 1087
Noterdaeme P. et al., 2012, A&A, 547, L1

Noterdaeme P., Petitjean P., Pâris I., Cai Z., Finley H., Ge J., Pieri M. M., York D. G., 2014, A&A, 566, A24
Padmanabhan H., Choudhury T. R., Refregier A., 2016, MNRAS, 458, 781
Pâris I. et al., 2017, A&A, 597, A79
Péroux C., McMahon R. G., Storrie-Lombardi L. J., Irwin M. J., 2003, MNRAS, 346, 1103
Planck Collaboration, 2016, A&A, 594, A13
Pontzen A. et al., 2008, MNRAS, 390, 1349
Prochaska J. X., Wolfe A. M., 1997, ApJ, 487, 73
Prochaska J. X., Wolfe A. M., 1998, ApJ, 507, 113
Prochaska J. X., Wolfe A. M., 2002, ApJ, 566, 68
Prochaska J. X., Wolfe A. M., 2009, ApJ, 696, 1543
Prochaska J. X., Gawiser E., Wolfe A. M., Castro S., Djorgovski S. G., 2003, ApJL, 595, L9
Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, ApJ, 635, 123
Rafelski M., Wolfe A. M., Prochaska J. X., Neeleman M., Mendez A. J., 2012, ApJ, 755, 89
Rahmati A., Schaye J., 2014, MNRAS, 438, 529
Ross N. P. et al., 2012, ApJS, 199, 3
Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
Slosar A. et al., 2011, JCAP, 9, 001
Smee S. A. et al., 2013, AJ, 146, 32
Tescari E., Viel M., Tornatore L., Borgani S., 2009, MNRAS, 397, 411
Tinker J. L., Robertson B. E., Kravtsov A. V., Klypin A., Warren M. S., Yepes G., Gottlöber S., 2010, ApJ, 724, 878
Villaescusa-Navarro F. et al., 2016, MNRAS, 456, 3553
Vladilo G., 2002, A&A, 391, 407
Vladilo G., Centurión M., Bonifacio P., Howk J. C., 2001, ApJ, 557, 1007
Wolfe A. M., Prochaska J. X., 1998, ApJL, 494, L15
Wolfe A. M., Prochaska J. X., 2000, ApJ, 545, 591
Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, ApJS, 61, 249
Wolfe A. M., Gawiser E., Prochaska J. X., 2005, ARA&A, 43, 861
Yèche C. et al., 2010, A&A, 523, A14
York D. G. et al., 2000, AJ, 120, 1579
Zafar T., Péroux C., Popping A., Milliard B., Deharveng J.-M., Frank S., 2013, A&A, 556, A141

APPENDIX A: PUBLIC ACCESS TO DATA AND CODE

The software used to generate the results in this paper is written in C^{++} and is publicly available at https://github.com/iprafols/cross_correlations. This repository also contains a PYTHON library with functions to plot the cross-correlation measurements, and the correlation function and configuration files necessary to reproduce our main results. Instructions to install and run the software are in the repository. Data and configuration files are in plain text format.

APPENDIX B: PROJECTOR OF THE δ FIELD AND THE DISTORTION MATRIX FORMALISM

B1 Motivation

As explained in Section 3, the assumption is that the measured Ly α transmission fluctuation, $\delta^{(m)}$, differs from the true Ly α transmission fluctuation, $\delta^{(t)}$ according to

$$\delta_i^{(m)} = \delta_i^{(t)} + a + b\lambda_i, \tag{B1}$$

where a and b are small unknown functions that depend on the δ field in a complicated manner, and λ is either the wavelength or the logarithm of the wavelength (whichever is used in the computation of the δ field). Here, we assume that a, b are constant within a given forest.

This hypothesis is motivated by the definition of the δ field. As explained in Section 2.2 the δ field is defined as

$$\delta_i = \frac{f_i}{C_q(\lambda_i)\,\overline{F}(z_i)} - 1, \tag{B2}$$

where f_i is the measured flux, $C_q(\lambda_i)$ is the quasar continuum (or unabsorbed flux) and $\overline{F}(z_i)$ is the mean transmitted fraction at the Ly α absorber redshift. The pixel redshift is $z_i = \lambda_i/\lambda_{\mathrm{Ly}\alpha} - 1$. The quasar continuum is assumed to have the form $C_q(\lambda_i) = \overline{C}(\lambda_i)(a + b\lambda_i)$, where $\overline{C}$ is the mean flux determined by stacking all quasar spectra, estimated at the restframe wavelength, and a and b are fitted constants, different for different forests. We can fit the parameters a and b except for a small error, i.e. $a = a_t - \delta_a$ and $b = b_t - \delta_b$.

If we Taylor expand this expression and retain only the leading order

$$\delta_i^{(m)} \approx \delta_i^{(t)} - \frac{\delta_a}{a_t + b_t\lambda_i} - \frac{\delta_b\lambda}{a_t + b_t\lambda_i}. \tag{B3}$$

We can now assume that the average of b_t along the different forests will be zero, and that for each individual forest, it is a small fluctuation of this average. This assumption is motivated by the fact that the steepness of the flux spectra is accounted for in the estimation of $\overline{C}$. Therefore, we can neglect $b_t\lambda$ over a, hence the presented hypothesis (equation B1).

B2 Projector

Since it is impossible to know the values of a and b in equation (B1), it is necessary to identify a projector, P, that allows the removal of these parameters, i.e.

$$P\delta^{(m)} = P\delta^{(t)}. \tag{B4}$$

To find an expression for this projector, it is useful to adopt a vectorial representation, which allows one to treat the forest as a whole. Keep in mind that we are assuming a and b to be constant throughout the forest.

To start with the derivation, we first consider the case $b = 0$. In vectorial form, and for a forest of length N, we have

$$\begin{pmatrix} \delta_1^{(m)} \\ \vdots \\ \delta_N^{(m)} \end{pmatrix} = \begin{pmatrix} \delta_1^{(t)} \\ \vdots \\ \delta_N^{(t)} \end{pmatrix} - a v_1, \tag{B5}$$

where v_1 is the vector $\frac{1}{\mathcal{N}_1}(1, \cdots, 1)$, and $\mathcal{N}_1$ is a normalization constant that makes the vector unitary, i.e. $v_1^t v_1 = 1$.

We now construct a projector P_0 that will cancel the second term in the equation above. This projector reads

$$P_0 = \mathbb{I} - v_1 v_1^t, \tag{B6}$$

and it does indeed cancel the second term in equation (B5):

$$P_0 v_1 = \left(\mathbb{I} - v_1 v_1^t\right) v_1 = \mathbb{I} v_1 - v_1 \underbrace{v_1^t v_1}_{=1} = v_1 - v_1 = 0. \tag{B7}$$

Now that we have an appropriate projector, let us relax the condition $b = 0$. We now have

$$\begin{pmatrix} \delta_1^{(m)} \\ \vdots \\ \delta_N^{(m)} \end{pmatrix} = \begin{pmatrix} \delta_1^{(t)} \\ \vdots \\ \delta_N^{(t)} \end{pmatrix} - a v_1 - b v_2, \tag{B8}$$

where $v_2 = (\lambda_1, \cdots, \lambda_N)$.

We must expand the projector P_0 to a new projector P in such a manner that maintains the condition $Pv_1 = 0$ imposed in the

particular case where $b = 0$ and add the extra condition that $Pv_2 = 0$, i.e. we have to project using a vector that is orthogonal to v_1. We can follow the Gram–Schmidt process to determine such a vector: $u_2 = v_2 - \left(v_2^t v_1\right) v_1$. Any vector in this direction will verify $P_1 v_1 = 0$. However, for it to verify $Pv_2 = 0$, we need a vector in the direction of u_2 that is properly normalized. Therefore, the new projector reads

$$P = \mathbb{I} - v_1 v_1^t - \frac{1}{\mathcal{N}_2^2} u_2 u_2^t, \tag{B9}$$

where $\mathcal{N}_2^2 = u_2^t u_2 = v_2^t v_2 - v_2^t v_1 v_1^t v_2$.

This projector verifies both conditions:

$$\begin{aligned} P v_1 &= \left[\mathbb{I} - v_1 v_1^t - \frac{1}{\mathcal{N}_2^2} u_2 u_2^t\right] v_1 \\ &= \underbrace{P_0 v_1}_{=0} - \frac{1}{\mathcal{N}_2^2} u_2 u_2^t v_1 = -\frac{1}{\mathcal{N}_2^2}\left[v_2 - \left(v_2^t v_1\right) v_1\right] \\ &\quad \times \left[v_2 - \left(v_2^t v_1\right) v_1\right]^t v_1 = -\frac{1}{\mathcal{N}_2^2}\left[v_2 - \left(v_2^t v_1\right) v_1\right] \\ &\quad \times \underbrace{\left(v_2^t v_1 - \left(v_2^t v_1\right) \underbrace{v_1^t v_1}_{=1}\right)}_{=0} = 0, \end{aligned} \tag{B10}$$

and

$$\begin{aligned} P v_2 &= \left[\mathbb{I} - v_1 v_1^t - \frac{1}{\mathcal{N}_2^2} u_2 u_2^t\right] v_2 = \mathbb{I} v_2 - v_1 v_1^t v_2 \\ &\quad - \frac{1}{\mathcal{N}_2^2}\left[v_2 - \left(v_2^t v_1\right) v_1\right]\left[v_2 - \left(v_2^t v_1\right) v_1\right]^t v_2 \\ &= v_2 - v_1 v_1^t v_2 - \frac{1}{\mathcal{N}_2^2}\left[v_2 - \left(v_2^t v_1\right) v_1\right] \\ &\quad \times \underbrace{\left(v_2^t v_2 - \left(v_2^t v_1\right) v_1^t v_2\right)}_{=\mathcal{N}^2} = v_2 - v_1\left(v_1^t v_2\right) \\ &\quad - \left(v_2 - v_1\left(v_1^t v_2\right)\right) = 0. \end{aligned} \tag{B11}$$

This projector allows us to compare the real and the measured values without knowing the parameters a and b. The derivation has been performed without specifying the scalar product, so the expression for the projector (equation B9) is then valid for any given scalar product. This behaviour is interesting because not all the pixels in the Ly α forest are equally noisy; there is a weight associated with each pixel. This weight is now easily introduced into this formalism if one simply defines the scalar product as

$$u^t v = \sum_{i \in f} u_i v_i w_i, \tag{B12}$$

where i is an index that runs over pixels in a particular forest f.

Now that we have specified the scalar product, we can find specific expressions for v_1, u_2 and $\mathcal{N}_2^2$.

$$\begin{aligned} (1 \cdots 1)\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} &= \sum_{i \in f} w_i \\ \Rightarrow v_1 &= \frac{1}{\sqrt{\sum_{i \in f} w_i}}(1, ..., 1), \end{aligned} \tag{B13}$$

$$\begin{aligned} u_2 &= v_2 - v_2^t v_1 v_1 = v_2 - \frac{\sum_{i \in f} \lambda_i w_i}{\sqrt{\sum_{i \in f} w_i}} v_1 \\ &= v_2 - \bar{\lambda}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 - \bar{\lambda} \\ \vdots \\ \lambda_N - \bar{\lambda} \end{pmatrix}, \end{aligned} \tag{B14}$$

and

$$\mathcal{N}_2^2 = u_2^t u_2 = \sum_{i \in f} \left(\lambda_i - \bar{\lambda}\right)^2 w_i, \tag{B15}$$

where $\bar{x} \equiv \sum_{i \in f} x w_i / \sum_{i \in f} w_i$.

Using this scalar product and the corresponding expressions for v_1, u_2 and $\mathcal{N}_2^2$ derived above, we can study the behaviour of this projector when it is applied to a vector δ, defined in the forest of interest.

$$\begin{aligned} P\delta &= \left[\mathbb{I} - v_1 v_1^t - \frac{1}{\mathcal{N}^2} u_2 u_2^t\right]\delta = \delta - v_1 v_1^t \delta \\ &\quad - \frac{1}{\mathcal{N}_2^2} u_2 u_2^t \delta = \delta - \frac{\sum_{i \in f} \delta_i w_i}{\sum_{i \in f} w_i}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ &\quad - \frac{\sum_{i \in f}\left(\lambda_i - \bar{\lambda}\right)\delta_i w_i}{\sum_{i \in f}\left(\lambda_i - \bar{\lambda}\right)^2 w_i}\begin{pmatrix} \lambda_1 - \bar{\lambda} \\ \vdots \\ \lambda_N - \bar{\lambda} \end{pmatrix}. \end{aligned} \tag{B16}$$

As we will see later, it is useful to consider the ith component of this vector:

$$\begin{aligned} (P\delta)_i &= \sum_{j \in f} P_{ij}\delta_j = \delta_i - \bar{\delta} - \frac{\sum_{j \in f} \delta_j\left(\lambda_j - \bar{\lambda}\right) w_j}{\sum_{j \in f}\left(\lambda_j - \bar{\lambda}\right)^2 w_j} \\ &\quad \times \left(\lambda_i - \bar{\lambda}\right). \end{aligned} \tag{B17}$$

B3 Distortion matrix

The χ^2 statistic for this estimator reads

$$\chi^2 = (\xi - \langle\xi\rangle)^t C^{-1}(\xi - \langle\xi\rangle), \tag{B18}$$

where C is the covariance matrix between the different bins of the cross-correlation.

The expected value of the cross-correlation estimator in bin A can be written as

$$\left\langle \xi^A \right\rangle = \frac{\sum_{d,f}\sum_{i \in f} \Theta_{id}^A w_i \sum_{j \in f} P_{ij}\xi_{jd}}{\sum_{d,f}\sum_{i \in f} \Theta_{id}^A w_i}, \tag{B19}$$

where the indexes d and f run over DLAs and forests, respectively, the indexes i and j run over pixels in a particular forest, Θ_{id}^A is 1 if the DLA-pixel pair is in bin A and 0 otherwise, and ξ_{jd} is the theoretical prediction of the cross-correlation for the DLA-pixel pair jd.

At this point, we can discretize the model similarly to the discretization of the data. Then

$$\xi_{jd} = \sum_B \Theta_{jd}^B \xi^B. \tag{B20}$$

This discretization need not be the same as the discretization on the data, but it is convenient to do so. The formalism presented here applies to whichever case is chosen.

Introducing this discretization into equation B19, the expected value of the cross-correlation can be written as

$$\left\langle \xi^A \right\rangle = \sum_B \frac{\sum_{d,f} \sum_{i\in f} \Theta^A_{id} w_i \sum_{j\in f} P_{ij} \Theta^B_{jd} \xi^B}{\sum_{d,f} \sum_{i\in f} \Theta^A_{id} w_i} \equiv \sum_B D^{AB} \xi^B, \tag{B21}$$

where D^{AB} is the distortion matrix element that relates the cross-correlation measured in bin A and the model for the cross-correlation in bin B. The quantity D^{AB} is defined as

$$D^{AB} = \frac{1}{\sum_{d,f} \sum_{i\in f} \Theta^A_{id} w_i} \sum_{d,f} \sum_{i\in f} \Theta^A_{id} w_i \times \sum_{j\in f} \left(\delta_{ij} - \frac{w_j}{\sum_{k\in f} w_k} - \frac{\left(\lambda_j - \bar{\lambda}_f\right)\left(\lambda_i - \bar{\lambda}_f\right) w_j}{\sum_{k\in f} \left(\lambda_k - \bar{\lambda}_f\right)^2 w_k} \right) \Theta^B_{jd}, \tag{B22}$$

where δ_{ij} is the Kronecker delta and should not be confused with the Ly α transmission fluctuation.

This paper has been typeset from a TEX/LATEX file prepared by the author.