Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **496,** 5436–5454 (2020) doi:10.1093/mnras/staa1806
Advance Access publication 2020 June 23

Detecting multiple DLAs per spectrum in SDSS DR12 with Gaussian processes

Ming-Feng Ho,[1]★ Simeon Bird[1] and Roman Garnett[2]

[1] *Department of Physics & Astronomy, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA*
[2] *Department of Computer Science and Engineering, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130, USA*

Accepted 2020 June 16. Received 2020 May 27; in original form 2020 April 1

ABSTRACT

We present a revised version of our automated technique using Gaussian processes (GPs) to detect damped Lyman α absorbers (DLAs) along quasar (QSO) sightlines. The main improvement is to allow our GP pipeline to detect multiple DLAs along a single sightline. Our DLA detections are regularized by an improved model for the absorption from the Lyman α forest that improves performance at high redshift. We also introduce a model for unresolved sub-DLAs that reduces misclassifications of absorbers without detectable damping wings. We compare our results to those of two different large-scale DLA catalogues and provide a catalogue of the processed results of our GP pipeline using 158 825 Lyman α spectra from SDSS data release 12. We present updated estimates for the statistical properties of DLAs, including the column density distribution function, line density (dN/dX), and neutral hydrogen density ($\Omega_{\rm DLA}$).

Key words: methods: statistical – intergalactic medium – quasars: absorption lines – galaxies: statistics.

1 INTRODUCTION

Damped Ly α absorbers (DLAs) are absorption line systems with high neutral hydrogen column densities ($N_{\rm H\,I} > 10^{20.3}$ cm^{-2}) discovered in sightlines of quasar (QSO) spectroscopic observations (Wolfe et al. 1986). The gas that gives rise to DLAs is dense enough to be self-shielded from the ultraviolet background (UVB; Cen 2012) yet diffuse enough to have a low star formation rate (Fumagalli et al. 2014). DLAs dominate the neutral-gas content of the Universe after reionization (Gardner et al. 1997; Noterdaeme et al. 2012; Zafar et al. 2013; Crighton et al. 2015). Simulations tell us DLAs are connected with galaxies over a wide range of halo masses (Prochaska & Wolfe 1997; Haehnelt, Steinmetz & Rauch 1998; Pontzen et al. 2008), and at $z \geq 2$ are formed from the accretion of neutral hydrogen gas on to dark matter haloes (Bird et al. 2014, 2015). The abundance of DLAs at different epochs of the Universe ($2 < z < 5$) thus becomes a powerful probe to understand the formation history of galaxies (Gardner et al. 1997; Wolfe, Gawiser & Prochaska 2005).

Finding DLAs historically involves a combination of template fitting and visual inspection of spectra by the eyes of trained astronomers (Prochaska, Herbert-Fort & Wolfe 2005; Slosar et al. 2011). Recent spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS) (York et al. 2000) have taken large amount of QSO spectra (Pâris et al. 2012, 2014) (∼500 000 in SDSS-IV, Pâris et al. 2018). Future surveys such as the Dark Energy Spectroscopic Instrument (DESI)[1] will acquire more than 1 million QSOs, making visual inspection of the spectra impractical. Moreover, the low signal-to-noise ratios of SDSS data make the task of detecting DLAs even harder, and induces noise related detection systematics. Since the release of the SDSS DR14 QSO catalogue (Pâris et al. 2018), visual inspection is no longer performed on all QSO targets. A fully automated and statistically consistent method thus needs to be presented for current and future surveys.

We provide a catalogue of DLAs using SDSS DR12 with 158 825 QSO sightlines. We demonstrate that our pipeline is capable of detecting an arbitrary number of DLAs within each spectroscopic observation, which makes it suitable for future surveys. Furthermore, since our pipeline resides within the framework of Bayesian probability, we have the ability to make probabilistic statements about those observations with low signal-to-noise ratios. This property allows us to make probabilistic estimations of DLA population statistics, even with low-quality noisy data (Bird, Garnett & Ho 2017).

Other available searches of DLAs in SDSS include: a visual inspection survey (Slosar et al. 2011), visually guided Voigt-profile fitting (Prochaska et al. 2005; Prochaska & Wolfe 2009); and three automated methods: a template-fitting method (Noterdaeme et al. 2012), an unpublished machine-learning approach using Fisher discriminant analysis (Carithers 2012), and a deep-learning approach using a convolutional neural network (Parks et al. 2018). Although these methods have had some success in creating large DLA catalogues, they suffer from hard-to-control systematics due to reliance either on templates or black box training.

We present a revised version of our previous automated method based on a Bayesian model-selection framework (Garnett et al. 2017). In our previous model (Garnett et al. 2017), we built a likelihood function for the QSO spectrum, including the continuum and the non-DLA absorption, using Gaussian processes (GPs; Rasmussen & Williams 2005). The SDSS DR9 concordance catalogue was applied to learn the covariance of the GP model. In this paper, we use the effective optical depth of the Lyman-series forest to allow the mean model of the likelihood function to be

★ E-mail: mho026@ucr.edu
[1] http://desi.lbl.gov

adjustable to the mean flux of the QSO spectrum, which reduces the probability of falsely fitting high column density absorbers at high redshifts. We also improve our knowledge of low column density absorbers and build an alternative model for sub-DLAs, which are the H I absorbers with $19.5 < \log_{10}N_{\rm H\,I} < 20$. These modifications allow us to extend our previous pipeline to detect an arbitrary number of DLAs within each QSO sightline without overfitting.

Alongside the revised DLA detection pipeline, we present the new estimates of DLA statistical properties at $z > 2$. Since the neutral hydrogen gas in DLAs will eventually accrete on to galactic haloes and fuel the star formation, these population statistics can give an independent constraint on the theory of galaxy formation. Our pipeline relies on a well-defined Bayesian framework and contains a full posterior density on the column density and redshift for a given DLA. We thus can properly propagate the uncertainty in the properties of each DLA spectrum to population statistics of the whole sample. Additionally, we are also able to account for low signal-to-noise ratio samples in our population statistics since the uncertainty will be reflected in the posterior probability. We thus substantially increase the sample size in our measurements by including these noisy observations.

2 NOTATION

We will briefly recap the notation we defined in Garnett et al. (2017). Imagine we are observing a QSO with a known redshift $z_{\rm QSO}$. The underlying true emission function $f(\lambda_{\rm rest})$ ($f : \mathcal{X} \rightarrow \mathbb{R}$) of the QSO is a mapping relation from rest-frame wavelength to flux. We will always assume the $z_{\rm QSO}$ is known and rescale the observed-frame wavelength $\lambda_{\rm obs}$ to the rest-frame wavelength with $\lambda_{\rm rest}(= \lambda_{\rm obs}/(1 + z_{\rm QSO}))$. We will use λ to replace $\lambda_{\rm rest}$ in the rest of the text because we only work on $\lambda_{\rm rest}$.

The QSO spectrum observed is not the intrinsic emission function $f(\lambda)$. Both the instrumental noise and absorption due to the intervening intergalactic medium along the line of sight will affect the observed flux. We thus denote the observed flux as a function $y(\lambda)$.

For a real spectroscopic observation, we measure the function $y(\lambda)$ on a discrete set of samples $\boldsymbol{\lambda}$. We thus denote the observed flux as a vector $\boldsymbol{y}$, which is defined as $y_i = y(\lambda_i)$ with i representing ith pixel. For a given QSO observation, we use $\mathcal{D}$ to represent a set of discrete observations $(\boldsymbol{\lambda}, \boldsymbol{y})$.

We exclude missing values of the spectroscopic observations in our calculations. These missing values are due to pixel masking in the spectroscopic observations (e.g. bad columns in the CCD detectors). We will use NaN ('not a number') to represent those missing values in the text, and we will always ignore NaNs in the calculations.

3 BAYESIAN MODEL SELECTION

The classification approach used in our pipeline depends on Bayesian model selection. Bayesian model selection allows us to compute the probability that a spectroscopic sightline $\mathcal{D}$ contains an arbitrary number of DLAs through evaluating the probabilities of a set of models $\{\mathcal{M}_i\}$, where i is a positive integer. This set of $\mathcal{M}_i$ contains all potential models we want to classify: a model with no DLA and models having between one DLA and k DLAs.

For each $\mathcal{M}_i$, we want to compute the probability that best explains the data $\mathcal{D}$ given a model $\mathcal{M}$. To do this, we have to marginalize the model parameters θ and evaluate the model evidence,

$$p(\mathcal{D} \mid \mathcal{M}) = \int p(\mathcal{D} \mid \mathcal{M}, \theta) p(\theta \mid \mathcal{M}) \mathrm{d}\theta. \tag{1}$$

Given a set of model pieces of evidence $p(\mathcal{D} \mid \mathcal{M}_i)$ and model priors $\Pr(\mathcal{M}_i)$, we are able to evaluate the posterior of a model given data based on Bayes's rule,

$$\Pr(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M}) \Pr(\mathcal{M})}{\sum_i p(\mathcal{D} \mid \mathcal{M}_i) \Pr(\mathcal{M}_i)}. \tag{2}$$

We will select the model from $\{M_i\}$ with the highest posterior. Readers may think of this method as an application of Bayesian hypothesis testing. Instead of only getting the likelihoods conditioned on models, we get posterior probabilities for each model given data.

Let k be the maximum number of DLAs we will want to detect in a QSO spectrum. For our multi-DLA model selection, we will develop $k + 2$ models, which include a null model for no DLA detection ($\mathcal{M}_{\neg \rm DLA}$), models for detecting exactly k DLAs ($\mathcal{M}_{\rm DLA(k)}$), and a model with sub-DLAs ($\mathcal{M}_{\rm sub}$). With a given spectroscopic sightline $\mathcal{D}$, we will compute the posterior probability of having exactly k DLAs in data $\mathcal{D}$, $\Pr(\mathcal{M}_{\rm DLA(k)} \mid \mathcal{D})$.

4 GAUSSIAN PROCESSES

In this section, we will briefly recap how we use GPs to describe the QSO emission function $f(\lambda)$, following Garnett et al. (2017). The QSO emission function is a complicated function without a simple form derived from physically motivated parameters. We thus use a non-parametric framework, GPs, for modelling this physically unknown function $f(\lambda)$. A detailed introduction to GPs may be found in Rasmussen & Williams (2005).

4.1 Definition and prior distribution

We wish to use a GP to model the QSO emission function $f(\lambda)$. We can treat a GP as an extension of the joint Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to infinite continuous domains. The difference is that a GP is a distribution over functions, not just a distribution over a finite number of random variables (although since we are dealing with pixelized variables here the distinction is less important).

A GP is completely specified by its first two central moments, a mean function $\mu(\lambda)$ and a covariance function $K(\lambda, \lambda')$:

$$\begin{aligned} \mu(\lambda) &= \mathbb{E}\left[f(\lambda) \mid \lambda\right], \\ K(\lambda, \lambda') &= \mathbb{E}\left[(f(\lambda) - \mu(\lambda))(f(\lambda') - \mu(\lambda')) \mid \lambda, \lambda'\right] \\ &= \mathrm{cov}\left[f(\lambda), f(\lambda') \mid \lambda, \lambda'\right]. \end{aligned} \tag{3}$$

The mean vector describes the expected behaviour of the function, and the covariance function specifies the covariance between pairs of random variables. We thus will write the GP as

$$f(\lambda) \sim \mathcal{GP}(\mu(\lambda), K(\lambda, \lambda')). \tag{4}$$

We can write the prior probability distribution of a GP as

$$p(f) = \mathcal{GP}(f; \mu, K). \tag{5}$$

Real spectroscopic observations measure a discrete set of inputs $\boldsymbol{\lambda}$ and the corresponding $f(\boldsymbol{\lambda})$, so we get a multivariate Gaussian distribution

$$p(\boldsymbol{f}) = \mathcal{N}(f(\boldsymbol{\lambda}); \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda}')). \tag{6}$$

Assuming the dimension of $\boldsymbol{\lambda}$ and $\boldsymbol{f}$ is d, the form of the multivariate Gaussian distribution is written as

$$\begin{aligned} &\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \mathbf{K}) \\ &= \frac{1}{\sqrt{(2\pi)^d \det \mathbf{K}}} \exp\left(-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{\mu})^\top \mathbf{K}^{-1} (\boldsymbol{f} - \boldsymbol{\mu})\right). \end{aligned} \tag{7}$$

4.2 Observation model

We now have a GP model for a discrete set of wavelengths $\boldsymbol{\lambda}$ and true emission fluxes $\boldsymbol{f}$. To build the likelihood function for observational data $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$, we have to incorporate the observational noise. Here, we assume the observational noise is modelled by an independent Gaussian variable for each wavelength pixel, allowing the noise realization to differ between pixels but neglecting inter-pixel correlations.

The noise variance for a given λ_i is written as $v_i = \sigma(\lambda_i)^2$. $\sigma(\lambda_i)$ is the measurement error from a single observation on a given wavelength point λ. With the above assumptions, we can write down the mechanism of generating observations as

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{v}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \mathbf{V}), \tag{8}$$

where $\mathbf{V} = \text{diag}\ \boldsymbol{v}$, which means we put the vector $\boldsymbol{v}$ on the diagonal terms of the diagonal square matrix $\mathbf{V}$.

Given an observational model $p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{v})$ and a GP emission model $p(\boldsymbol{f} \mid \boldsymbol{\lambda})$, the prior distribution for observations $\boldsymbol{y}$ is obtained by marginalizing the latent function $\boldsymbol{f}$:

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}) &= \int p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{v}) p(\boldsymbol{f} \mid \boldsymbol{\lambda}) \mathrm{d}\boldsymbol{f} \\ &= \int \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \mathbf{V}) \mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \mathbf{K}) \mathrm{d}\boldsymbol{f} \\ &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{K} + \mathbf{V}), \end{aligned} \tag{9}$$

where the Gaussians are closed under the convolution. Our observation model thus becomes a multivariate normal distribution described by a mean model $\mu(\boldsymbol{\lambda})$, covariance structure $K(\lambda, \lambda')$, and the instrumental noise $\mathbf{V}$. The instrumental noise is derived from SDSS pipeline noise, so it is different from QSO-to-QSO; however, since $\mathbf{K}$ encodes the covariance structure of QSO emissions, $\mathbf{K}$ should be the same for all QSOs.

As explained in Garnett et al. (2017), there is no obvious choice for a prior covariance function $\mathbf{K}$ for modelling the QSO emission function. Most off-the-shelf covariance functions assume some sort of translation invariance, but this is not suitable for spectroscopic observations.[2] However, we understand the QSO emission function will be independent of the presence of a low-redshift DLA. We also assume that QSO emission functions are roughly redshift independent in the wavelength range of interest (Lyman limit to Lyman α), as accretion physics should not strongly vary with cosmological evolution. We thus build our own custom μ and K for the GP prior to model the QSO spectra.

5 LEARNING A GP PRIOR FROM QSO SPECTRA

In this section, we will recap the prior modelling choices we made in Garnett et al. (2017) and the modifications we made to reliably detect multiple DLAs in one spectrum. We first build a GP model for QSO emission in the absence of DLAs, the null model $\mathcal{M}_{\neg\text{DLA}}$. Our model with DLAs ($\mathcal{M}_{\text{DLA}}$) extends this null model. With the model priors and model evidence of all models we are considering, we compute the model posterior with Bayesian model selection.

The GP prior is completely described by the first two moments, the mean and covariance functions, which we derive from data. We must consider the mean flux of QSO emission, the absorption effect due to the Lyman α forest, and the covariance structure within the Lyman series.

5.1 Data

Our training set to learn our GP null model comprises the spectra observed by SDSS BOSS DR9 and labelled as containing (or not) a DLA by Lee et al. (2013).[3] The DR9 data set includes 54 468 QSO spectra with $z_{\text{QSO}} > 2.15$. We removed the following QSOs from the training set:

(i) $z_{\text{QSO}} < 2.15$: QSOs with redshifts lower than 2.15 have no Lyman α in the SDSS band.

(ii) `BAL`: QSOs with broad absorption lines as flagged by the SDSS pipeline.

(iii) Spectra with less than 200 detected pixels.

(iv) `ZWARNING`: spectra whose analysis had warnings as flagged by the SDSS redshift estimation. Extremely noisy spectra (the `TOO_MANY_OUTLIERS` flag) were kept.

5.2 Modelling decisions

Consider a set of QSO observations $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$; we always shift the observer's frame λ_{obs} to rest-frame λ so that we can set the emissions of Lyman series from different spectra to the same rest wavelengths. The assumption here is that the z_{QSO}s of QSOs are known for all the observed spectra, which is not precisely true for the spectroscopic data we have here. Accurately estimating the redshift of QSOs is beyond the scope of this paper, and is tackled elsewhere (Fauber et al. 2020).

The observed magnitude of a QSO varies considerably, based on its luminosity distance and the properties of the black hole. For the observation $\boldsymbol{y}$ to be described by a GP, it is necessary to normalize all flux measurements by dividing by the median flux observed between 1310 and 1325 Å, a wavelength region that is unaffected by the Lyman α forest.

We model the same wavelength range as in Garnett et al. (2017):

$$\lambda \in [911.75\text{Å}, 1215.75\text{Å}], \tag{10}$$

going from the QSO rest-frame Lyman limit to the QSO rest-frame Lyman α. The spacing between pixels is $\Delta\lambda = 0.25$ Å. Note that we prefer not to include the region past the Lyman limit. This is partly due to the relatively small amount of data in that region and partly because the non-Gaussian Lyman break associated with Lyman limit systems can confuse the model. In particular, it occasionally tries to model a Lyman break with a wide DLA profile with a high column density. We shall see this is especially a problem if the QSO redshift is slightly inaccurate. The code considers the prior probability of a Lyman break at a higher redshift than the putative QSO rest frame to be zero and thus is especially prone to finding other explanations for the large absorption trough.

To model the relationship between flux measurements and the true QSO emission spectrum, we have to add terms corresponding to instrumental noise and weak Lyman α absorption to the intrinsic correlations within the emission spectrum. Instrumental noise was already added in equation (9) as a matrix $\mathbf{V}$.

The remaining part of the modelling is to define the GP covariance structure for QSOs across different redshifts. In Garnett et al. (2017), Lyman α absorbers were modelled by a single additive noise term, $\boldsymbol{\Omega}$,

[2]Detailed explanations are in Garnett et al. (2017), section 4.2.1.

[3]However, we use the DR12 pipeline throughout.

accounting for the effect of the forest as extra noise in the emission spectrum. This is not completely physical: it assumes that the Lyman α forest is just as likely to cause emission as absorption.

Here, we rectify this by not only including the Lyman α perturbation term in our GP as $\mathbf{\Omega}$, but introducing a redshift-dependent mean flux ($\mu(z)$) with a dependence on the absorber redshift ($z(\lambda_{\rm obs})$). We model the overall mean model with a redshift-dependent absorption function and a mean emission vector: $\mu(z) = a(z) \circ \boldsymbol{\mu}$. The notation $\circ$ refers to Hadamard product, which is the element-wise product between two vectors or matrices. The covariance matrix is decomposed into $\mathbf{A}_{\rm F}(\mathbf{K} + \mathbf{\Omega})\mathbf{A}_{\rm F}$, where diag $(\mathbf{A}_{\rm F}) = a(z)$ and $\mathbf{A}_{\rm F}$ is a diagonal matrix.[4] The $\mathbf{K}$ matrix describes the covariance between different emission lines in the QSO spectrum, which we will learn from data. The $\mathbf{A}_{\rm F}$ matrix is applied to $\mathbf{K}$ because we assume that $\mathbf{K}$ is learned before the absorption noise $a(z)$ is applied. See Section 5.4 for how we learn the covariance.

Combining all modelling decisions, the model prior for an observed QSO emission is

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, \mathcal{M}_{\neg \rm DLA}) = \mathcal{N}(\boldsymbol{y}; \mu(z), \mathbf{A}_{\rm F}(\mathbf{K} + \mathbf{\Omega})\mathbf{A}_{\rm F} + \mathbf{V}). \tag{11}$$

The mean emission flux is now redshift- and wavelength-dependent, so the optimization steps will differ slightly from Garnett et al. (2017). We will address the modifications in the following subsections.

5.3 Redshift-dependent mean flux vector

In this paper, instead of using a single mean vector $\boldsymbol{\mu}$ to describe all spectra, we adjust the mean model of the GP to fit the mean flux of each QSO spectrum. For modelling the effect of forest absorption on the flux, we adopt an empirical power law with effective optical depth $\tau_0(1+z)^\beta$ for Ly α forest (Kim et al. 2007):

$$a(z) = \exp\left(-\tau_0(1+z)^\beta\right), \tag{12}$$

where the absorber redshift z is related to the observer's wavelength $\lambda_{\rm obs}$ as

$$\begin{aligned} 1 + z &= \frac{\lambda_{\rm obs}}{\lambda_{{\rm Ly}\alpha}} \\ &= \frac{\lambda_{\rm obs}}{1215.7\text{Å}} \\ &= (1 + z_{\rm QSO})\frac{\lambda}{1215.7\text{Å}}, \end{aligned} \tag{13}$$

so the absorber redshift $z(\lambda_{\rm obs}) = z(\lambda, z_{\rm QSO})$ is a function of the QSO redshift and the wavelength.

In Garnett et al. (2017), we assumed the absorption from the forest would only play a role in the additive noise term ($\boldsymbol{\omega}$) in our likelihood model $p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\omega}, z_{\rm QSO}, \mathcal{M}_{\neg \rm DLA})$ with the form:

$$\omega'(\lambda, \lambda_{\rm obs}) = \omega(\lambda) s(z(\lambda_{\rm obs}))^2; \tag{14}$$

$$s(z) = 1 - \exp\left(-\tau_0(1+z)^\beta\right) + c_0, \tag{15}$$

where z is the absorber redshift. The $\omega(\lambda)$ term represents the global absorption noise, and the $s(z)$ corresponds to the absorption effect contributed by the Lyman α absorbers along the line of sight as a function of the absorber redshift z.

Thus in our earlier model, the Lyman α forest introduces additional fluctuations in the observed spectrum $\boldsymbol{y}$. This assumption worked well for low-redshift spectra, because mean absorption due to the Lyman α forest at low redshifts is relatively small. At high redshifts, however, the suppression of the mean flux induced by many Lyman α absorbers is substantial, see Fig. 1. In our earlier model, essentially all high-redshift QSO spectra were substantially more absorbed than the mean emission model μ due to absorption from the Lyman α forest. To explain this absorption, our model would fit multiple DLAs with large column densities.

We have improved the modelling of the Lyman α forest by allowing the mean GP model μ to be redshift dependent, having a mean optical depth following the measurement of Kim et al. (2007):

$$\begin{aligned} \tau_{\rm eff}(z) &= \tau_0(1+z)^\gamma \\ &= 0.0023 \times \exp(1+z)^{3.65}. \end{aligned} \tag{16}$$

There are other measurements of $\tau_{\rm eff}$ at higher precision than Kim et al. (2007; e.g. Becker et al. 2013). However, they are derived from SDSS data while Kim et al. (2007) were derived from high-resolution spectra. We therefore choose to use Kim et al. (2007) to preserve the likelihood principle that priors should not depend on the data set in question.

We include the effect of the whole Lyman series with a similar model, but however accounting for the different atomic coefficients of the higher order Lyman lines:

$$\begin{aligned} &\tau_{\rm eff,H\,I}(z(\lambda_{\rm obs}); \gamma, \tau_0) \\ &= \sum_{i=2}^{N} \tau_0 \frac{\lambda_{1i} f_{1i}}{\lambda_{12} f_{12}} (1 + z_{1i}(\lambda_{\rm obs}))^\gamma \times I_{(z_{1i}(\min(\lambda_{\rm obs})), z_{\rm QSO})}(z). \end{aligned} \tag{17}$$

Here, f_{1i} represents the oscillator strength and λ_{1i} corresponds to the transition wavelength from the $n = 1$ to $n = i$ atomic energy level. We model the Lyman series up to $N = 32$, with $i = 2$ being Ly α and $i = 3$ Ly β. The absorption redshift z_{1i} for the $n = 1$ to $n = i$ transition is defined by

$$1 + z_{1i} = \frac{\lambda_{\rm obs}}{\lambda_{1i}}. \tag{18}$$

The optical depth at the line centre is estimated by

$$\tau_0 = \sqrt{\pi} \frac{e^2}{m_e c} \frac{N_\ell f_{\ell u} \lambda_{\ell u}}{b}, \tag{19}$$

where ℓ indicates the lower energy level and u is the upper energy level. For Lyman α, we have $\lambda_{\ell u} = 1215.7$ Å and $f_{\ell u} = 0.4164$; for Lyman β, we have $\lambda_{\ell u} = 1025.7$ and $f_{\ell u} = 0.07912$. Given equation (19), we have the effective optical depth for the Lyman β forest:

$$\tau_\beta = \frac{f_{31}\lambda_{31}}{f_{21}\lambda_{21}}\tau_0 = \frac{0.07912 \times 1025.7}{0.4164 \times 1215.7} \times 0.0023 = 0.0004. \tag{20}$$

The mean prior of the GP model for each spectrum is rewritten as

$$\boldsymbol{\mu}(z) = \boldsymbol{\mu} \circ \exp\left(-\tau_{\rm eff,H\,I}(z; \gamma = 3.65, \tau_0 = 0.0023)\right). \tag{21}$$

We will simply write $\tau_{\rm eff,H\,I}(z) = \tau_{\rm eff,H\,I}(z; \gamma = 3.65, \tau_0 = 0.0023)$ in the following text for simplicity. The new $\boldsymbol{\mu}$ is estimated via

$$\boldsymbol{\mu} = \frac{1}{N_{\neg \rm NaN}} \sum_{y_{ij} \neq \rm NaN} y_{ij} \cdot \exp\left(+\tau_{\rm eff,H\,I}(z_{ij})\right). \tag{22}$$

Equation (22) rescales the mean observed fluxes back to the expected continuum before the suppression due to Lyman series absorption, hopefully recovering approximately the true QSO emission function $\boldsymbol{f}$. Fig. 1 shows the retrained mean QSO emission model for an example QSO. The mean model, $\boldsymbol{\mu}$, is much closer to the peak emission flux above the absorbed forest.

[4] $\mathbf{A}_{\rm F}^{\rm T} = \mathbf{A}_{\rm F}$ because it is diagonal.

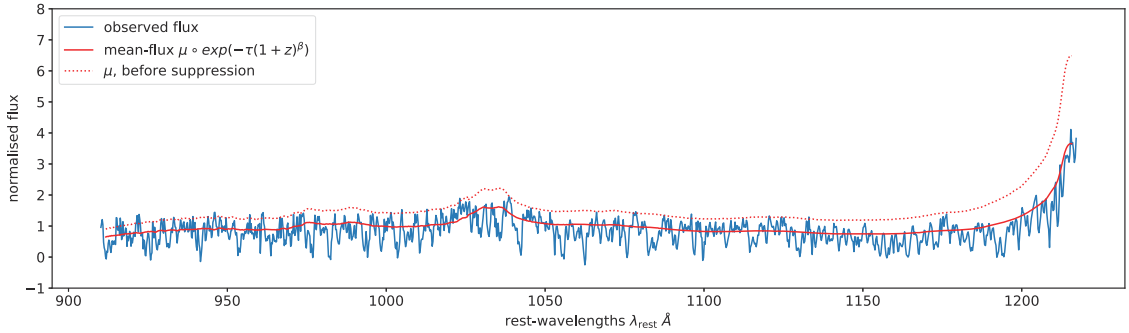

Figure 1. The effect of the shift to the GP mean vector from the Lyman α forest effective optical depth model ($\boldsymbol{\mu} \circ \exp(-\tau_0(1+z)^{\beta})$). The dotted red curve shows the mean emission model before application of the forest suppression. The solid red curve is the mean model including the forest suppression.

For model consistency, we account for the mean suppression from weak absorbers in our redshift-dependent noise model ω with:

$$\omega'(\lambda, \lambda_{\rm obs}) = \omega(\lambda) s_F(z(\lambda_{\rm obs}))^2; \tag{23}$$

$$\text{where } s_F(z(\lambda_{\rm obs})) = 1 - \exp(-\tau_{\rm eff,H\,I}(z(\lambda_{\rm obs}); \beta, \tau_0)) + c_0\,. \tag{24}$$

τ_0, β, and c_0 are parameters that are learned from the data. Fig. 2 shows the mean model and absorption noise variance we use, compared to the model from Garnett et al. (2017).

Note that the mean flux model introduces degeneracies between the parameters of equation (24). For example, c_0 may be compensated by the overall amplitude of pixel-wise noise vector $\boldsymbol{\omega}$. For this reason, we should not ascribe strict physical interpretations to the optimal values of equation (24). The optimized $\boldsymbol{\omega}'$ is simply an empirical relation modelling the pixel-wise and redshift-dependent noise in the null model given SDSS data.

After introducing the effective optical depth into our GP mean model, we decrease the number of large DLAs we detect at high redshifts and thus measure lower $\Omega_{\rm DLA}$ at high redshifts (see Section 10.3 for more details). This is because, for high-redshift QSOs, the mean optical depth may be close to unity. To explain this unexpected absorption, the previous code will fit multiple high column density absorbers to the raw emission model, artificially increasing the number of DLAs detected. With the mean model suppressed, there is substantially less raw absorption to explain, and so this tendency is avoided.

5.4 Learning the flux covariance

$\mathbf{K}$ and Ω (equation 11) are optimized to maximize the likelihood of generating the data, $\mathcal{D}$. The mean flux model is not optimized, but follows the effective optical depth reported in Kim et al. (2007). Thus, we remove the effect of forest absorption before we train the covariance function and train on $\mathcal{D}' = \{\boldsymbol{\lambda}, \boldsymbol{y} \circ \exp(+\tau_{\rm eff,H\,I}(z)) - \mu(z)\}$ to find the optimal parameters for $\mathbf{K}$ and Ω.

We assume the same likelihood as Garnett et al. (2017) for generating the whole training data set ($\boldsymbol{Y}$):

$$p(\mathbf{Y} \mid \boldsymbol{\lambda}, \mathbf{V}, \mathbf{M}, \boldsymbol{\omega}, z_{\rm QSO}, \mathcal{M}_{\neg\rm DLA}) = \prod_{i=1}^{N_{\rm spec}} \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}_i), \tag{25}$$

where $\boldsymbol{Y}$ means the matrix containing all the observed flux in the training data, and the product on the right-hand side says we are combining all likelihoods from each single spectrum. The noise matrix $\boldsymbol{\Omega} = \operatorname{diag} \boldsymbol{\omega}'$ is the diagonal matrix that represents the Lyman α forest absorption from equation (24).

$\mathbf{M}$ is a low-rank decomposition of the covariance matrix $\mathbf{K}$ we want to learn:

$$\mathbf{K} = \mathbf{M}\mathbf{M}^{\top}, \tag{26}$$

where $\mathbf{M}$ is an $(N_{\rm pixels} \times k)$ matrix. Without this low-rank decomposition, we would need to learn $N^2_{\rm pixel} = 1217 \times 1217$ free parameters. With equation (26), we can limit the number of free parameters to be $N_{\rm pixels} \times k$, where $k \ll N_{\rm pixels}$; also, it guarantees the covariance matrix $\mathbf{K}$ to be positive semidefinite. Each column of the $\mathbf{M}$ can be treated as an eigenspectrum of the training data, where we set the number of eigenspectra to be $k = 20$. We will optimize the $\mathbf{M}$ matrix and the absorption noise in equation (24) simultaneously.

A modification performed in this work is to, instead of directly training on the observed flux, optimize the covariance matrix and noise model on the flux with Lyman α forest absorption removed (de-forest flux):

$$\begin{aligned} \boldsymbol{y} &:= \boldsymbol{y} \circ \exp(+\tau_{\rm eff,H\,I}(z)); \\ Y_{ij} &:= Y_{ij} \exp(+\tau_{\rm eff,H\,I}(z))_{ij}\,. \end{aligned} \tag{27}$$

We may write this change into the likelihood:

$$p(\mathbf{Y} \circ \exp(+\tau_{\rm eff,H\,I}(z)) \mid \boldsymbol{\lambda}, \mathbf{V}, \mathbf{M}, \boldsymbol{\omega}, z_{\rm QSO}, \mathcal{M}_{\neg\rm DLA}) = \prod_{i=1}^{N_{\rm spec}} \mathcal{N}(\boldsymbol{y}_i \circ \exp(+\tau_{\rm eff,H\,I}(z_i)); \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}_i), \tag{28}$$

where $\boldsymbol{\mu}$ is the mean model from equation (22). The rest of our optimization procedure follows the unconstrained optimization of Garnett et al. (2017).

We use de-forest fluxes for training as we want our covariance matrix to learn the covariance in the true emission function. The emission function (like our kernel $\mathbf{K}$) is independent of QSO emission redshift, whereas the absorption noise is not. We only implement the mean forest absorption of Kim et al. (2007), so we need an extra term to compensate for the variance of the forest around this mean. We thus still train the redshift- and wavelength-dependent absorption noise from data. The optimal values we learned for equation (24) are:

$$c_0 = 0.3050; \tau_0 = 1.6400 \times 10^{-4}; \beta = 5.2714. \tag{29}$$

As we might expect, the optimal τ_0 value is smaller than the $\tau_0 = 0.01178$ learned in Garnett et al. (2017), which implies the effect of the forest is almost removed by applying the Lyman-series forest to the mean model. The final covariance matrix is shown in Fig. 3.

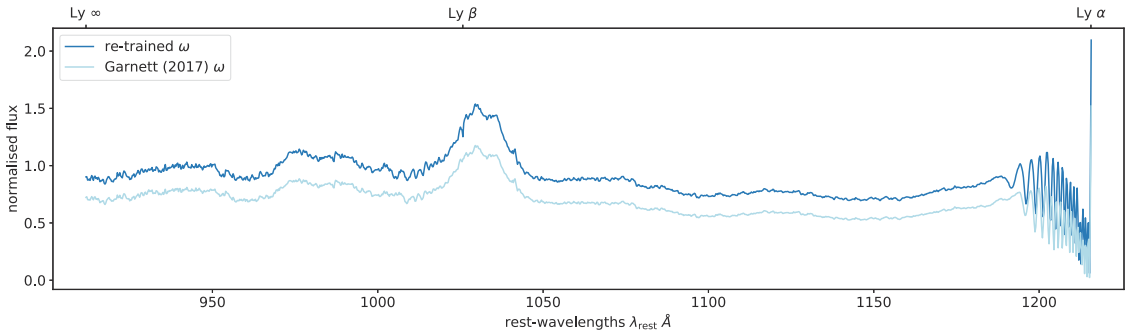

Figure 2. The difference between the original pixel-wise noise variance $\boldsymbol{\omega}$ (Garnett et al. 2017) and the retrained $\boldsymbol{\omega}$ from equation (28). The retrained $\boldsymbol{\omega}$ decreases because the fit no longer needs to account for the mean forest absorption.

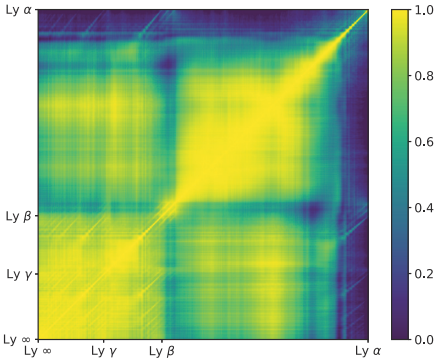

Figure 3. The trained covariance matrix $\mathbf{M}$, which is almost the same as the covariance from Garnett et al. (2017). Note that we normalize the diagonal elements to be unity, so this is more like a correlation matrix than a covariance matrix. The values in the matrix are ranging from 0 to 1, representing the correlation between λ and λ' in the QSO emission.

5.5 Model evidence

Consider a given QSO observation $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$ with known observational noise $\nu(\boldsymbol{\lambda})$ and known QSO redshift $z_{\rm QSO}$. The model evidence for $\mathcal{M}_{\neg\rm DLA}$ can be estimated using

$$p(\mathcal{D} \mid \mathcal{M}_{\neg\rm DLA}, \nu, z_{\rm QSO}) \propto p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, \mathcal{M}_{\neg\rm DLA}), \tag{30}$$

which is equivalent to evaluating a multivariate Gaussian

$$\begin{aligned} &p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, \mathcal{M}_{\neg\rm DLA}) \\ &\quad = \mathcal{N}(\boldsymbol{y}; \mu \circ \exp(-\boldsymbol{\tau}_{\rm eff,H\,I}), \mathbf{A}_{\rm F}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\rm F} + \mathbf{V}). \end{aligned} \tag{31}$$

Here, $\exp(-\boldsymbol{\tau}_{\rm eff,H\,I}) = \text{diag}\,\mathbf{A}_{\rm F}$ describes the absorption due to the forest and modifies the mean vector $\boldsymbol{\mu}$, the covariance matrix $\mathbf{K}$, and the noise matrix $\boldsymbol{\Omega}$ to account for the Lyman α forest effective optical depth.

6 A GP MODEL FOR QSO SIGHTLINES WITH MULTIPLE DLAS

In Section 5, we learned a GP prior for QSO spectroscopic measurements without any DLAs for our null model $\mathcal{M}_{\neg\rm DLA}$. Here, we extend the null model $\mathcal{M}_{\neg\rm DLA}$ to a model with k intervening DLAs, $\mathcal{M}_{\rm DLA(k)}$.

Our complete DLA model, $\mathcal{M}_{\rm DLA}$, will be the union of the models with i DLAs: $\mathcal{M}_{\rm DLA} = \{\mathcal{M}_{\rm DLA(}i\rm)}\}_{i=1}^{k}$. We consider only until $k = 4$, as DLAs are rare events and our sample only contains one spectrum with four DLAs.

6.1 Absorption function

Before we model a QSO spectrum with intervening DLAs, we need to have an absorption profile model for a DLA. Damped Lyman alpha absorbers, or DLAs, are neutral hydrogen (H I) absorption systems with saturated lines and damping wings in the spectroscopic measurements. Having saturated lines means the column density of the absorbers on the line of sight is high enough to absorb essentially all photons. The damping wings are due to natural broadening in the line.

The optical depth from each Lyman series transition is

$$\tau(\lambda; z_{\rm DLA}, N_{\rm H\,I}) = N_{\rm H\,I}\frac{\pi e^2 f_{1u}\lambda_{1u}}{m_e c}\phi(v, b, \gamma), \tag{32}$$

where e is the elementary charge, λ_{1u} is the transition wavelength from the $n = 1$ to $n = u$ energy level ($\lambda_{12} = 1215.6701$ Å for Lyman α), and f_{1u} is the oscillator strength of the transition. The line profile ϕ is a Voigt profile:

$$\begin{aligned} \phi(v, b, \gamma) = &\int \frac{\mathrm{d}v}{\sqrt{2\pi}\sigma_v}\exp(-v^2/2\sigma_v^2) \\ &\times \frac{4\gamma_{\ell u}}{16\pi^2[v - (1 - v/c)\nu_{\ell u}]^2 + \gamma_{\ell u}^2}, \end{aligned} \tag{33}$$

which is a convolution between a Lorenztian line profile and a Gaussian line profile. The σ_v is the one-dimensional velocity dispersion, $\gamma_{\ell u}$ is a parameter for Lorenztian profile, ν is the frequency, and u represents the upper energy level and ℓ represents the lower energy level.

Both profiles are parametrized by the relative velocity v, which means both profiles are distributions in the one-dimensional velocity space:

$$v = c\left(\frac{\lambda}{\lambda_{1u}}\frac{1}{(1 + z_{\rm DLA})} - 1\right). \tag{34}$$

The standard deviation of the Gaussian line profile is related to the broadening parameter $b = \sqrt{2}\sigma_v$, and if we assume the broadening is entirely due to thermal motion:

$$b = \sqrt{\frac{2kT}{m_p}}. \tag{35}$$

Introducing the damping constant $\Gamma = 6.265 \times 10^8 \mathrm{s}^{-1}$ for Lyman α, we have the parameter $\gamma_{\ell u}$ to describe the width of the Lorenztian profile

$$\gamma_{\ell u} = \frac{\Gamma\lambda_{\ell u}}{4\pi}. \tag{36}$$

Our default DLA profile includes Ly α, Ly β, and Ly γ absorptions. We fix the broadening parameter b by setting $T = 10^4$ K, which increases the width of the DLA profile by 13 km s^{-1}, small compared to the effect of the Lorenztian wings. Thus, for a given QSO and a true emission function $f(\lambda)$, the function for the observed flux $y(\lambda)$ is

$$y(\lambda) = f(\lambda) \exp\left(-\tau(\lambda; z_{\rm DLA}, N_{\rm H\,I})\right) \exp\left(-\tau_{\rm eff,H\,I}(\lambda_{\rm obs})\right) + \epsilon, \quad (37)$$

where ϵ is additive Gaussian noise including measurement noise and absorption noise.

Suppose we have a DLA at redshift $z_{\rm DLA}$ with column density $N_{\rm H\,I}$. We can model the spectrum with an intervening DLA by calculating the DLA absorption function:

$$\boldsymbol{a} = \exp\left(-\tau(\boldsymbol{\lambda}; z_{\rm DLA}, N_{\rm H\,I})\right). \quad (38)$$

We apply the absorption function to the GP prior of $\boldsymbol{y}$ with

$$\begin{aligned} &p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, z_{\rm DLA}, N_{\rm H\,I}, \mathcal{M}_{\rm DLA}) \\ &\quad = \mathcal{N}(\boldsymbol{y}; \boldsymbol{a} \circ (\boldsymbol{a}_{\rm F} \circ \boldsymbol{\mu}), \mathbf{A}(\mathbf{A}_{\rm F}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\rm F})\mathbf{A} + \mathbf{V}), \end{aligned} \quad (39)$$

where $\mathbf{A} = \text{diag}\, \boldsymbol{a}$.

For a model with k DLAs with $k \in \mathbb{N}$, we simply take the element-wise product of k absorption functions:

$$\begin{aligned} \boldsymbol{a}_{(k)} &= \prod_{i=1}^{k} a(\boldsymbol{\lambda}; z_{{\rm DLA}\,i}, N_{{\rm H\,I}\,i}); \\ \text{diag}\, \mathbf{A}_{(k)} &= \boldsymbol{a}_{(k)}. \end{aligned} \quad (40)$$

The prior for $\mathcal{M}_{\rm DLA(k)}$ would therefore be

$$\begin{aligned} &p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, \{z_{{\rm DLA}\,i}\}_{i=1}^{k}, \{N_{{\rm H\,I}\,i}\}_{i=1}^{k}, \mathcal{M}_{\rm DLA(k)}) \\ &\quad = \mathcal{N}(\boldsymbol{y}; \boldsymbol{a}_{(k)} \circ (\boldsymbol{a}_{\rm F} \circ \boldsymbol{\mu}), \mathbf{A}_{(k)}(\mathbf{A}_{\rm F}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\rm F})\mathbf{A}_{(k)} + \mathbf{V}). \end{aligned} \quad (41)$$

Here, we briefly review our notations in equation (41): $\boldsymbol{a}_{(k)}$, which is parametrized by $(\{z_{{\rm DLA}\,i}\}_{i=1}^{k}, \{N_{{\rm H\,I}\,i}\}_{i=1}^{k})$, represents the absorption function with k DLAs in one spectrum. Note that each DLA is parametrized by a pair of $(z_{\rm DLA}, N_{\rm H\,I})$. $\boldsymbol{a}_{\rm F}$ corresponds to the absorption function from the Lyman series absorptions, which is derived from Kim et al. (2007) in the form of equation (21). The covariance matrix $\mathbf{K}$ and the absorption model $\boldsymbol{\Omega}$ are both learned from data, as described in Section 5.4. $\mathbf{V}$ is the noise variance matrix given by the SDSS pipeline, so each sightline would have different $\mathbf{V}$.

6.2 Model evidence: DLA(1)

The model evidence of our DLA model is given by the integral:

$$\begin{aligned} p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(1)}, z_{\rm QSO}) \propto \int & p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, \theta, z_{\rm QSO} \mathcal{M}_{\rm DLA(1)}) \\ & \times p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}) {\rm d}\theta, \end{aligned} \quad (42)$$

where we integrated out the parameters, $\theta = (z_{\rm DLA}, \log_{10} N_{\rm H\,I})$, with a given parameter prior $p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)})$.

However, equation (42) is intractable, so we approximate it with a quasi-Monte Carlo method (QMC). QMC selects $N = 10\,000$ samples with an approximately uniform spatial distribution from a Halton sequence to calculate the model likelihood, approximating the model evidence by the sample mean:

$$p(D \mid \mathcal{M}_{\rm DLA(1)}, z_{\rm QSO}) \simeq \frac{1}{N} \sum_{i=1}^{N} p(\mathcal{D} \mid \theta_i, z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}). \quad (43)$$

6.3 Model evidence: Occam's razor effect for DLA(k)

For higher order DLA models, we have to integrate out not only the nuisance parameters of the first DLA model $\mathcal{M}_{\rm DLA(1)}$, (θ_1) but also the parameters from $\mathcal{M}_{\rm DLA(2)}$ to $\mathcal{M}_{\rm DLA(k)}$,

$$\begin{aligned} p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(k)}, z_{\rm QSO}) \propto \int & p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(k)}, \{\theta_i\}_{i=1}^{k}) \\ & \times p(\{\theta_i\}_{i=1}^{k} \mid \mathcal{M}_{\rm DLA(k)}, \mathcal{D}, z_{\rm QSO}) {\rm d}\{\theta_i\}_{i=1}^{k}, \end{aligned} \quad (44)$$

which means we are marginalizing $\{\theta_i\}_{i=1}^{k}$ in a parameter space with $2 \times k$ dimensions. The parameter prior of multi-DLAs is a multiplication between a non-informative prior $p(\theta_i \mid \mathcal{M}_{\rm DLA(1)}, z_{\rm QSO})$ and the posterior of the $(k-1)$ multi-DLA model,

$$\begin{aligned} &p(\{\theta_i\}_{i=1}^{k} \mid \mathcal{M}_{\rm DLA(k)}, \mathcal{D}, z_{\rm QSO}) \\ &\quad = p(\{\theta_i\}_{i=1}^{k-1} \mid \mathcal{M}_{\rm DLA(k-1)}, \mathcal{D}, z_{\rm QSO}) p(\theta_k \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}). \end{aligned} \quad (45)$$

We can approximate this integral using the same QMC method. For example, if we want to sample the model evidence for $\mathcal{M}_{\rm DLA(2)}$, we would need $N = 10\,000$ samples for each parameter dimension $\{\theta_i\}_{i=1}^{2}$, which results in sampling from two independent Halton sequences with 10^8 samples in total. If we want to sample up to $\mathcal{M}_{\rm DLA(k)}$ with N samples for each $\{\theta_i\}$ from $i = 1, ..., k$, we would need to have:

$$\begin{aligned} p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(k)}, z_{\rm QSO}) \simeq & \frac{1}{N} \sum_{j^{(1)}=1}^{N} \frac{1}{N} \sum_{j^{(2)}=1}^{N} \frac{1}{N} \sum_{j^{(3)}=1}^{N} \cdots \frac{1}{N} \sum_{j^{(k)}=1}^{N} \\ & p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(k)}, \{\theta_{1j^{(1)}}, \theta_{2j^{(2)}}, \theta_{3j^{(3)}}, \ldots, \theta_{kj^{(k)}}\}, z_{\rm QSO}), \end{aligned} \quad (46)$$

where $\{j^{(1)}, j^{(2)}, j^{(3)}, \ldots, j^{(k)}\}$ indicate the indices of QMC samples. The above equation (46) is thus in principle evaluated with N^k samples.

In practice, we only sample $N = 10\,000$ points from $p(\{\theta_i\}_{i=1}^{k} \mid \mathcal{M}_{\rm DLA(k)}, \mathcal{D}, z_{\rm QSO})$ instead of sampling N^k points, as a uniform sampling of the first DLA model may be reweighted to cover parameter space for the higher order models. A N^{k-1} factor of normalization is thus left behind in the summation,

$$\begin{aligned} &p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(k)}, z_{\rm QSO}) \\ &\simeq \frac{1}{N^k} \sum_{j=1}^{N} p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(k)}, \{\theta_{ij}\}_{i=1}^{k}, z_{\rm QSO}) \\ &\simeq \frac{1}{N^{k-1}} \left(\frac{1}{N} \sum_{j=1}^{N} p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(k)}, \{\theta_{ij}\}_{i=1}^{k}, z_{\rm QSO}) \right) \\ &\simeq \frac{1}{N^{k-1}} \text{mean}_j \left(p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(k)}, \{\theta_{ij}\}_{i=1}^{k}, z_{\rm QSO}) \right). \end{aligned} \quad (47)$$

The additional $\frac{1}{N^{k-1}}$ factor penalizes models with more parameters than needed, and can be viewed as an implementation of Occam's razor. This Occam's razor effect is caused by the fact that all probability distributions have to be normalized to unity. A model with more parameters, which means having a wider distribution in the likelihood space, results in a bigger normalization factor.

The motivation for us to draw N samples from the multi-DLA likelihood function $p(\{\theta_i\}_{i=1}^{k} \mid \mathcal{M}_{\rm DLA(k)}, \mathcal{D}, z_{\rm QSO})$ is that we believe the prior density we took from the posterior density of $\mathcal{M}_{\rm DLA(k-1)}$ is representative enough even without N^k samples. For example, if we have two peaks in our likelihood density $p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(1)}, \theta_1, z_{\rm QSO})$, we expect the sampling for θ_2 in $p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(2)}, \{\theta_1, \theta_2\}, z_{\rm QSO})$

would concentrate on sampling the density of the first highest peak in $p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(1)}, \theta_1, z_{\rm QSO})$ density. Similarly, while we are sampling for $\mathcal{M}_{\rm DLA(3)}$, we expect θ_3 and θ_2 would cover the first- and the second-highest peaks.

To avoid multi-DLAs overlapping with each other, we inject a dependence between any pair of $z_{\rm DLA}$ parameters. Specifically, if any pair of $z_{\rm DLA}$s have a relative velocity smaller than 3000 km s^{-1}, then we set the likelihood of this sample to NaN.

6.4 Additional penalty for DLAs and sub-DLAs

In Section 6.3, we apply a penalty, Occam's razor, to regularize DLA models using more parameters than needed. This effect is due to the normalization (to unity) of the evidence.

In a similar fashion, and for a similar reason, we apply an additional regularization factor between the non-DLA and DLA models (including sub-DLAs). This additional factor ensures that when both models are a poor fit to a particular observational spectrum, the code prefers the non-DLA model, rather than preferring the model with more parameters and thus greater fitting freedom. We directly inject this Occam's razor factor in the model selection:

$$\Pr(\mathcal{M}_{\rm DLA} \mid \mathcal{D}) = \frac{\Pr(\mathcal{M}_{\rm DLA}) p(\mathcal{D} \mid \mathcal{M}_{\rm DLA}) \frac{1}{N}}{\left(\begin{array}{l} \Pr(\mathcal{M}_{\rm DLA}) p(\mathcal{D} \mid \mathcal{M}_{\rm DLA}) \\ +\Pr(\mathcal{M}_{\rm sub}) p(\mathcal{D} \mid \mathcal{M}_{\rm sub}) \end{array} \right) \frac{1}{N} + \Pr(\mathcal{M}_{\neg \rm DLA} \mid \mathcal{D})}, \quad (48)$$

where $N = 10^4$ is the number of samples we used to approximate the parametrized likelihood functions. We evaluated the impact of this regularization factor on the area under the curve (AUC) in the receiver-operating characteristics (ROC) plot.[5] For $N = 10^4$, the AUC changed from 0.949 to 0.960. We considered other penalty values and found that the AUC increased up to $N = 10^4$ and then plateaued.

In addition, we found by examining specific examples that this penalty regularized a relatively common incorrect DLA detection: finding objects in short, very noisy low redshift ($z \sim 2.2$) spectra. In these spectra our earlier model would prefer the DLA model purely because of its large parameter freedom. In particular a high column density DLA, large enough that the damping wings exceed the width of the spectrum, would be preferred. Such a fit exploits a degeneracy in the model between the mean observed flux and the DLA column density when the spectrum is shorter than the putative DLA. The Occam's razor penalty avoids these spurious fits by penalizing the extra parametric freedom in the DLA model.

6.5 Parameter prior

Here, we briefly recap the priors on model parameters chosen in Garnett et al. (2017). Suppose we want to make an inference for the column density and redshift of an absorber $\theta = (N_{\rm H\,I}, z_{\rm DLA})$ from a given spectroscopic observation, the joint density for the parameter prior would be

$$p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}) = p(N_{\rm H\,I}, z_{\rm DLA} \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}). \quad (49)$$

Suppose the absorber redshift and the column density are conditionally independent and the column density is independent of the QSO redshift $z_{\rm QSO}$:

$$p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}) = p(z_{\rm DLA} \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}) \times p(N_{\rm H\,I} \mid \mathcal{M}_{\rm DLA(1)}) \quad (50)$$

[5]See Section 10.1 for how we compute our ROC plot.

We set a bounded uniform prior density for the absorber redshift $z_{\rm DLA}$:

$$p(z_{\rm DLA} \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}) = \mathcal{U}[z_{\rm min}, z_{\rm max}], \quad (51)$$

where we define the finite prior range to be

$$z_{\rm min} = \max \left\{ \begin{array}{c} \frac{\lambda_{\rm Ly\,\infty}}{\lambda_{\rm Ly\,\alpha}} (1 + z_{\rm QSO}) - 1 + 3000 \text{ km s}^{-1}/c \\ \frac{\min \lambda_{\rm obs}}{\lambda_{\rm Ly\,\alpha}} - 1 \end{array} \right. \quad (52)$$

$$z_{\rm max} = z_{\rm QSO} - 3000 \text{ km s}^{-1}/c; \quad (53)$$

which means we have a prior belief that the centre of the absorber is within the observed wavelengths. The range of observed wavelengths is either from Ly ∞ to Ly α of the QSO rest frame ($\lambda_{\rm rest} \in [911.75$ Å, 1216.75 Å]) or from the minimum observed wavelength to Ly α. We also apply a conservative cut-off of 3000 km s^{-1} near to Ly ∞ and Ly α. The -3000 km s^{-1} cut-off for $z_{\rm max}$ helps to avoid proximity ionization effects due to the QSO radiation field. Furthermore, the $+3000$ km s^{-1} cutoff for $z_{\rm min}$ avoids a potentially incorrect measurement for $z_{\rm QSO}$. An underestimated $z_{\rm QSO}$ can produce a Lyman-limit trough within the region of the QSO expected to contain only Lyman-series absorption, and the code can incorrectly interpret this as a DLA.

For the column density prior, we follow Garnett et al. (2017). We first estimate the density of DLAs column density $p(N_{\rm H\,I} \mid \mathcal{M}_{\rm DLA})$ using the BOSS DR9 Lyman α forest sample. We choose to put our prior on the base-10 logarithm of the column density $\log_{10} N_{\rm H\,I}$ due to the large dynamic range of DLA column densities in SDSS DR9 samples.

We thus estimate the density of logarithm column densities $p(\log_{10} N_{\rm H\,I} \mid \mathcal{M}_{\rm DLA(1)})$ using univariate Gaussian kernels on the reported $\log_{10} N_{\rm H\,I}$ values in DR9 samples. Column densities from DLAs in DR9 with $N_{\rm DLA} = 5\,854$ are used to non-parametrically estimate the logarithm $N_{\rm H\,I}$ prior density, with

$$p_{\rm KDE}(\log_{10} N_{\rm H\,I} \mid \mathcal{M}_{\rm DLA(1)}) = \frac{1}{N_{\rm DLA}} \sum_{i=1}^{N_{\rm DLA}} \mathcal{N}(\log_{10} N_{\rm H\,I}; l_i, \sigma^2), \quad (54)$$

where l_i is the logarithm column density $\log_{10} N_{\rm H\,I}$ of the ith sample. The bandwidth σ^2 is selected to be the optimal value for a normal distribution, which is the default setting for MATLAB.

We further simplify the non-parametric estimate into a parametric form with

$$p_{\rm KDE}(\log_{10} N_{\rm H\,I} = N \mid \mathcal{M}_{\rm DLA(1)}) \simeq q(\log_{10} N_{\rm H\,I} = N) \propto \exp(aN^2 + bN + c); \quad (55)$$

where the parameters (a, b, c) for the quadratic function are fitted via standard least-squared fitting to the non-parametric estimate of density $p_{\rm KDE}(\log_{10} N_{\rm H\,I} \mid \mathcal{M}_{\rm DLA(1)})$ with the range $\log_{10} N_{\rm H\,I} \in [20, 22]$. The optimal values for the quadratic terms were

$$a = -1.2695; b = 50.863; c = -509.33. \quad (56)$$

Note that we have the same values as in Garnett et al. (2017).

Finally, we choose to be conservative about the data-driven column density prior. We thus take a mixture of a non-informative lognormal prior with the data-driven prior to make a non-restrictive prior on a large dynamical range:

$$p(\log_{10} N_{\rm H\,I} \mid \mathcal{M}_{\rm DLA(1)}) = \alpha q(\log_{10} N_{\rm H\,I} = N) + (1 - \alpha) \mathcal{U}[20, 23]. \quad (57)$$

Here we choose the mixture coefficient $\alpha = 0.97$, which favours the data-driven prior. We still include a small component of a non-

informative prior so that we are able to detect DLAs with a larger column density than in the training set, if any are present in the larger DR12 sample. Note that $\alpha = 0.97$ is 7 per cent higher than the coefficient chosen in Garnett et al. (2017), which was $\alpha = 0.90$. Our previous prior slightly overestimated the number of very large DLAs.

6.6 Sub-DLA parameter prior

As reported in Bird et al. (2017), the column density distribution function (CDDF) exhibited an edge feature: an overdetection of DLAs at low column densities ($\sim 10^{20}\,\mathrm{cm}^{-2}$). This did not affect the statistical properties of DLAs as we restrict column density to $N_{\mathrm{H\,I}} \geqslant 10^{20.3}\,\mathrm{cm}^{-2}$ for both line densities ($\mathrm{d}N/\mathrm{d}X$) and total column densities (Ω_{DLA}). However, to make our method more robust, here we describe a complementary method to avoid overestimating the number of low column density absorbers.

The excess of DLAs at $\sim 10^{20}\,\mathrm{cm}^{-2}$ is due to our model excluding lower column density absorbers such as sub-DLAs. Since we limited our column density prior of DLAs to be larger than $10^{20}\,\mathrm{cm}^{-2}$, the code cannot correctly classify a sub-DLA. Instead it correctly notes that a sub-DLA spectrum is more likely to be a DLA with a minimal column density than an unabsorbed spectrum.

To resolve our ignorance, we introduce an alternative model $\mathcal{M}_{\mathrm{sub}}$ to account the model posterior of those low column density absorbers in our Bayesian model selection. The likelihood function we used for sub-DLAs is identical to the one we built for DLA model $\mathcal{M}_{\mathrm{DLA}(1)}$ in equation (39) but has a different parameter prior on the column densities $p(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{sub}})$. We restricted our prior belief of sub-DLAs to be within the range $\log_{10} N_{\mathrm{H\,I}} \in [19.5, 20]$, and, as we do not have a catalogue of sub-DLAs for learning the prior density, we put a uniform prior on $\log_{10} N_{\mathrm{H\,I}}$:

$$p(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{sub}}) = \mathcal{U}[19.5, 20]. \tag{58}$$

We place a lower cut-off at $\log_{10} N_{\mathrm{H\,I}} = 19.5$ because the relatively noisy SDSS data offer limited evidence for absorbers with column densities lower than this limit.

7 MODEL PRIORS

Bayesian model selection allows us to combine prior information with evidence from the data-driven model to obtain a posterior belief about the detection of DLAs $p(\mathcal{M}_{\mathrm{DLA}} \mid \mathcal{D})$ using Bayes' rule. For a given spectroscopic observation $\mathcal{D}$, we already have the ability to compute the model evidence for a DLA ($p(\mathcal{D} \mid \mathcal{M}_{\mathrm{DLA}})$) and no DLA ($p(\mathcal{D} \mid \mathcal{M}_{\neg\mathrm{DLA}})$). However, to compute the model posteriors, we need to specify our prior beliefs in these models. Here, we approximate our prior belief $\Pr(\mathcal{M}_{\mathrm{DLA}})$ using the SDSS DR9 DLA catalogue.

Consider a QSO observation $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$ at z_{QSO}. We want to find our prior belief that $\mathcal{D}$ contains a DLA. We count the fraction of QSO sightlines in the training set containing DLAs with redshift less than $z_{\mathrm{QSO}} + z'$, where $z' = 30\,000\,\mathrm{km\,s}^{-1}/c$ is a small constant. If N is the number of QSO sightlines with redshift less than $z_{\mathrm{QSO}} + z'$, and M is the number of sightlines in this set containing DLAs in the QSO rest-frame wavelengths range we search, then our empirical prior for $\mathcal{M}_{\mathrm{DLA}}$ is

$$\Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}) = \frac{M}{N}. \tag{59}$$

We can break down our DLA prior $\Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}})$ for multiple DLAs in a QSO sightline $\Pr(\mathcal{M}_{\mathrm{DLA(k)}} \mid z_{\mathrm{QSO}})$ via

$$\Pr(\mathcal{M}_{\mathrm{DLA(k)}} \mid z_{\mathrm{QSO}}) \simeq \left(\frac{M}{N}\right)^{k} - \left(\frac{M}{N}\right)^{k+1}. \tag{60}$$

For example, $\frac{M}{N}$ represents our prior belief of having at least one DLA in the sightline, and $(\frac{M}{N})^2$ represents having at least two DLAs. $\frac{M}{N} - (\frac{M}{N})^2$ is thus our prior belief of having exactly one DLA at the sightline.

7.1 Sub-DLA model prior

The CDDF of Bird et al. (2017) exhibited an edge effect at $\log_{10} N_{\mathrm{H\,I}} \sim 20$ due to a lack of sampling at lower column densities. We thus construct an alternative model for lower column density absorbers (sub-DLAs, DLAs' lower column density cousins) to regularize DLA detections. We use the same GP likelihood function as the DLA model $\mathcal{M}_{\mathrm{DLA}}$ to compute our sub-DLA model evidence $p(\mathcal{D} \mid \mathcal{M}_{\mathrm{sub}})$ but with a different column density prior $p(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{sub}})$.

There is no sub-DLA catalogue available for us to estimate the empirical prior directly. We, therefore, approximate our sub-DLA model prior by rescaling our DLA model prior:

$$\Pr(\mathcal{M}_{\mathrm{sub}} \mid z_{\mathrm{QSO}}) \propto \Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}), \tag{61}$$

and we require our prior beliefs to sum to unity:

$$\Pr(\mathcal{M}_{\neg\mathrm{DLA}} \mid z_{\mathrm{QSO}}) + \Pr(\mathcal{M}_{\mathrm{sub}} \mid z_{\mathrm{QSO}}) + \Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}) = 1. \tag{62}$$

The scaling factor between the DLA prior and sub-DLA prior should depend on our prior probability density of the column density of the absorbers. Here, we assume the density of sub-DLA $\log_{10} N_{\mathrm{H\,I}}$ is an uniform density with a finite range of $\log_{10} N_{\mathrm{H\,I}} \in [19.5, 20]$. We believe there are more sub-DLAs than DLAs as high column density systems are generally rarer. We thus assume the probability of finding sub-DLAs at a given $\log_{10} N_{\mathrm{H\,I}}$ is the same as the probability of finding DLAs at the most probable $\log_{10} N_{\mathrm{H\,I}}$, which is

$$\begin{aligned} &p(\log_{10} N_{\mathrm{H\,I}} = N \mid \{\mathcal{M}_{\mathrm{DLA}}, \mathcal{M}_{\mathrm{sub}}\}) \\ &\quad = \alpha q(N \mid \mathcal{M}_{\mathrm{DLA}}) \mathbb{I}_{(20,23)}(N) \\ &\quad + \alpha \max\left(q(N \mid \mathcal{M}_{\mathrm{DLA}})\right) \mathbb{I}_{(19.5,20)}(N) \\ &\quad + (1-\alpha)\mathcal{U}[19.5, 23]. \end{aligned} \tag{63}$$

Since $q(N \mid \mathcal{M}_{\mathrm{DLA}})$ has a simple quadratic functional form, we can solve the maximum value analytically, which is $\max\left(q(N \mid \mathcal{M}_{\mathrm{DLA}})\right) \simeq q(N = 20.03 \mid \mathcal{M}_{\mathrm{DLA}})$.

We thus can use our prior knowledge about the logarithm of column densities for different absorbers to rescale model priors:

$$\Pr(\mathcal{M}_{\mathrm{sub}} \mid z_{\mathrm{QSO}}) = \frac{Z_{\mathrm{sub}}}{Z_{\mathrm{DLA}}} \Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}), \tag{64}$$

where the scaling factor is

$$\frac{Z_{\mathrm{sub}}}{Z_{\mathrm{DLA}}} = \frac{\int_{19.5}^{20} p(N \mid \{\mathcal{M}_{\mathrm{DLA}}, \mathcal{M}_{\mathrm{sub}}\}) \mathrm{d}N}{\int_{20}^{23} p(N \mid \{\mathcal{M}_{\mathrm{DLA}}, \mathcal{M}_{\mathrm{sub}}\}) \mathrm{d}N}, \tag{65}$$

which is the odds of finding absorbers in the range of $\log_{10} N_{\mathrm{H\,I}} \in [19.5, 20]$ compared to finding absorbers in $\log_{10} N_{\mathrm{H\,I}} \in [20, 23]$. Note that we will treat the model posteriors of the sub-DLA model as part of the non-detections of DLAs in the following analysis sections.

8 CATALOGUE

The original parameter prior in Garnett et al. (2017) is uniformly distributed in $z_{\rm DLA}$ between the Lyman limit ($\lambda_{\rm rest} = 911.76$ Å) and the Ly α emission of the QSO. In Bird et al. (2017), we chose the minimum value of $z_{\rm DLA}$ to be at the Ly β emission line of the QSO rest frame (instead of the Lyman limit) to avoid the region containing unmodelled Ly β forest. The primary reason for this was that the original absorption noise model did not include Ly β absorption. With the updated model from equation (24) we are able to model this absorption. Hence, for our new public catalogue, we sample $z_{\rm DLA}$ to be from Ly∞ to Lyα in the QSO rest frame and for the convenience of future investigators our public catalogue contains DLAs throughout the whole available spectrum, including Ly β to Ly ∞. There is still some contamination in the blue end of high-redshift spectra from the Ly β forest and occasional Lyman breaks from a misestimated QSO redshift. In practice we shall see that the contamination is not severe except for $z_{\rm DLA} > 3.75$. However, in the interest of obtaining as reliable DLA statistics as possible, when computing population statistics we consider only 3000 Å redward of Ly β to 3000 Å blueward of Ly α in the QSO rest frame.

In this paper, we computed the posterior probability of $\mathcal{M}_{\neg\rm DLA}$ to $\mathcal{M}_{\rm DLA(k)}$ models. For each spectrum, the catalogue includes:

(i) The range of redshift DLA searched $[z_{\rm min}, z_{\rm max}]$,
(ii) The log model priors from $\log \Pr(\mathcal{M}_{\neg\rm DLA} \mid z_{\rm QSO})$, $\log \Pr(\mathcal{M}_{\rm sub} \mid z_{\rm QSO})$, to $\log \Pr(\{\mathcal{M}_{{\rm DLA}(i)}\}_{i=1}^{k} \mid z_{\rm QSO})$,
(iii) The log model evidence $\log p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, \mathcal{M})$, for each model we considered,
(iv) The model posterior $\Pr(\mathcal{M} \mid \mathcal{D}, z_{\rm QSO})$, for each model we considered,
(v) The probability of having DLAs $\Pr(\{\mathcal{M}_{\rm DLA}\} \mid \mathcal{D}, z_{\rm QSO})$,
(vi) The probability of having zero DLAs $\Pr(\mathcal{M}_{\neg\rm DLA} \mid \mathcal{D}, z_{\rm QSO})$,
(vii) The sample log likelihoods $\log p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\rm QSO}, \{z_{{\rm DLA}(i)}\}_{i=1}^{k}, \{\log_{10} N_{{\rm H\,I}(i)}\}_{i=1}^{k}, \mathcal{M}_{{\rm DLA}(k)})$ for all DLA models we considered, and
(viii) The maximum a posteriori (MAP) values of all DLA models we considered.

The full catalogue will be available alongside the paper: http://tiny.cc/multidla_catalog_gp_dr12q. The code to reproduce the entire catalogue will be posted in https://github.com/rmgarnett/gp_dla_detection/tree/master/multi_dlas.

8.1 Running time

We ran our multi-DLA code on UCR's High-Performance Computing Center (HPCC) and Amazon Elastic Compute Cloud (EC2). The computation of model posteriors of $\mathcal{M}_{\neg\rm DLA}$, $\mathcal{M}_{\rm sub}$, $\{\mathcal{M}_{{\rm DLA}(i)}\}_{i=1}^{4}$ takes 7–11 s per spectrum on a 32-core node in HPCC and 3–5 seconds on a 48-core machine in EC2. For each spectrum, we have to compute $10\,000 * 5 + 1$ log likelihoods in the form of equation (11). If we scale the sample size from $N = 10\,000$ to $100\,000$, it costs 38–52 s on a 32-core node in HPCC.

9 EXAMPLE SPECTRA

Here, we show a few examples of the fitted GP priors, both to compare our method to others and to aid the reader in understanding concretely how our method works.

We show an exmple where our older model detected only one DLA, as shown in Fig. 4, while our new code detects three DLAs in this single spectrum as shown in Fig. 5. Because the mean QSO model includes a redshift-dependent term corresponding to intervening absorbers, our new mean model can now fit the mean observed QSO spectrum better. Although we show the sample likelihoods in the $\mathcal{M}_{\rm DLA(1)}$ parameter space, our current code finds these three DLAs in the six-dimensional parameter space $(z_{{\rm DLA}(i)}, \log_{10} N_{{\rm H\,I}(i)})_{i=1}^{3}$.

In Fig. 6, we show a representative sample of a very common case in our $\mathcal{M}_{\rm DLA(1)}$ model. The red curve represents our GP prior on the given spectrum, and the orange curve is the curve with fitted DLAs provided by the CNN model presented in Parks et al. (2018).[6] We found Parks et al. (2018) underestimated the column densities of the underlying DLAs in the spectra due to not modelling Lyman β and Lyman γ absorption in DLAs, while the predictions of $N_{\rm H\,I}$ in our model are more robust since the predicted $N_{\rm H\,I}$ is constrained by α, β, and γ absorption. In the spectrum, Lyman β absorption is clearly visible (although noisy). In Fig. 6, Parks et al. (2018) have actually mistaken the Ly γ absorption line of the DLA for another, weaker, DLA. This demonstrates again the necessity of including other Lyman-series members in the modelling steps. Since Parks et al. (2018) broke down each spectrum into pieces during the training and testing phases, it is impossible for the CNN to use knowledge about other Lyman series lines associated with the DLAs. Another example, from a spectrum where we detect two DLAs and the CNN detects four (although at low significance) is shown in Fig. 7. Here, the CNN has mistaken both the Ly β and Ly γ absorption associated with the large DLA at $z \sim 3$ (near the QSO rest frame) for separate DLAs at $z = 2.4$ and $z = 2.22$, respectively. The large DLA at $z \sim 3$ has been split into two of reduced column density and reduced confidence. The CNN has also missed the second genuine DLA at a rest-frame wavelength of 1025 Å, presumably due to the proximity of an emission line. Our code, able to model the higher order Lyman lines, has used the information contained within them to correctly classify this spectrum as containing two DLAs.

Fig. 9 shows an example that was problematic in both the models of Garnett et al. (2017) and Parks et al. (2018). This is an extremely noisy spectrum, where the length of the spectrum is not long enough for us to contain higher order Lyman-series absorption or even to see the full length of the putative Lyman α absorption. By eye, distinguishing a DLA from the noise is challenging. If we examine the sample likelihoods from our model (shown in Fig. 10), we see that the DLA posterior probability is spread over the whole of parameter space; in other words, all models are a poor fit for this noise-dominated spectrum. The model selection is thus really comparing the likelihood function on the basis of how much parametric freedom it has. After implementing the additional Occam's razor factor between the null model and parametrized models (DLAs and sub-DLAs) described in Section 6.4, we found that the large DLA fitted to the noisy short spectrum by Garnett et al. (2017) was no longer preferred. This indicates that our Occam's razor penalty is effective. As shown in Fig. 16, $\Omega_{\rm DLA}$ at low redshifts is lower than the measurements in Bird et al. (2017), indicating that this class of error is common enough to have a measurable effect on the column density function. We checked that the addition of the Occam's razor penalty, $\Omega_{\rm DLA}$ is insensitive to the noise threshold used when selecting the spectra for our sample.

There are still some very high redshift QSOs ($z_{\rm QSO} \gtrsim 5$) where our code clearly detects too many DLAs in a single spectrum, even at low redshift. We exclude these spectra from our population statistics. At

[6] We used the version of Parks et al. (2018)'s catalogue listed in the published paper and found on Google Drive at https://tinyurl.com/cnn-dlas.

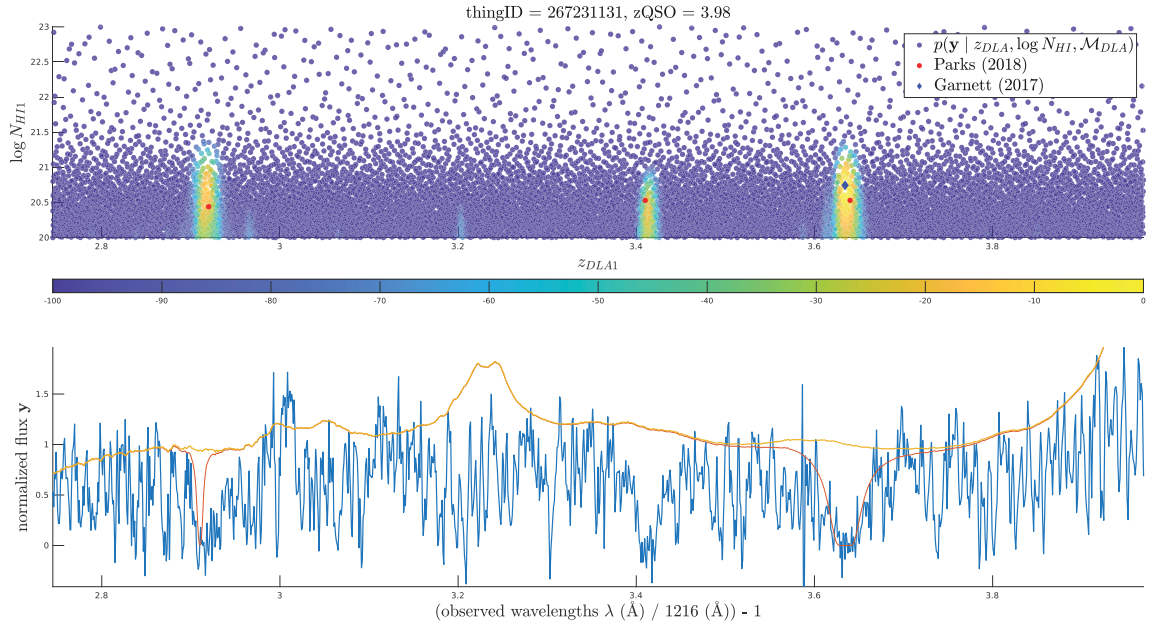

Figure 4. An example of finding DLAs using Garnett et al. (2017)'s model. Here, we use the single-DLA per spectrum version of Garnett's model. Upper: sample likelihoods $p(\boldsymbol{y} \mid \theta, \mathcal{M}_{\rm DLA})$ in the parameter space $\theta = (z_{\rm DLA}, \log_{10} N_{\rm H\,I})$. Red dots show the DLAs predicted by Parks et al. (2018), and the blue squares show the MAP prediction of the Garnett et al. (2017). Bottom: the observed spectrum (blue), the null model GP prior (orange), and the DLA model GP prior (red). So that the upper and bottom panels have the same *x*-axis, we rescale the observed wavelength to absorber redshift.

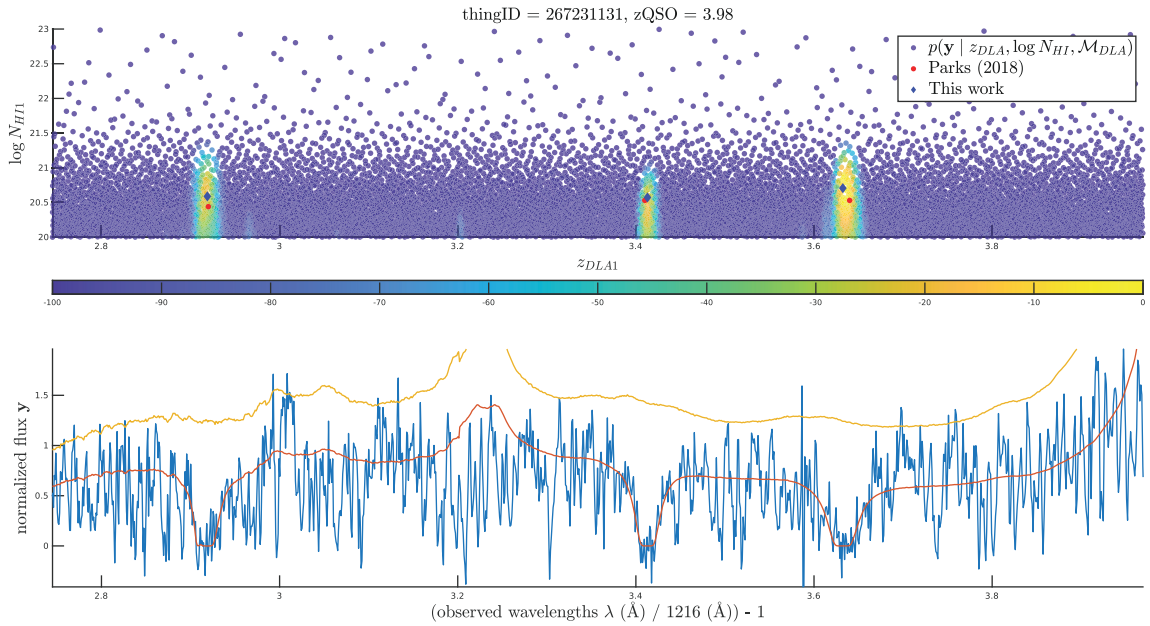

Figure 5. The same spectrum as Fig. 4, but using the multi-DLA model reported in this paper. Upper: sample likelihoods $p(\boldsymbol{y} \mid \theta, \mathcal{M}_{\rm DLA})$ in the parameter space of the $\mathcal{M}_{\rm DLA(1)}$, with $\theta = (z_{\rm DLA}, \log_{10} N_{\rm H\,I})$. Bottom: the observed spectrum (blue), the null model GP prior before the suppression of effective optical depth (orange), and the multi-DLA GP prior (red). The orange curve is slightly higher than the one in Fig. 4 because we try to model the mean spectrum before the forest. However, the DLA QSO model (red curve) matches the level of the observed mean flux better than Fig. 4 due to the inclusion of a term for the effective optical depth of the Lyman α forest.

high redshift, the Lyman α forest absorption is so strong as to render the observed flux close to zero. We thus cannot easily distinguish between the null model and the DLA models. It is also possible that at high redshifts, the mean flux of the forest is substantially different from the Kim et al. (2007) model we assume, and that this biases the fit. Finally, there are few such spectra, and so we cannot rule out the possibility that covariance of their emission spectra differs quantitatively from lower redshift QSOs.

10 ANALYSIS OF THE RESULTS

In this section, we present results from our classification pipeline, and we also present the statistical properties (CDDF, line densities

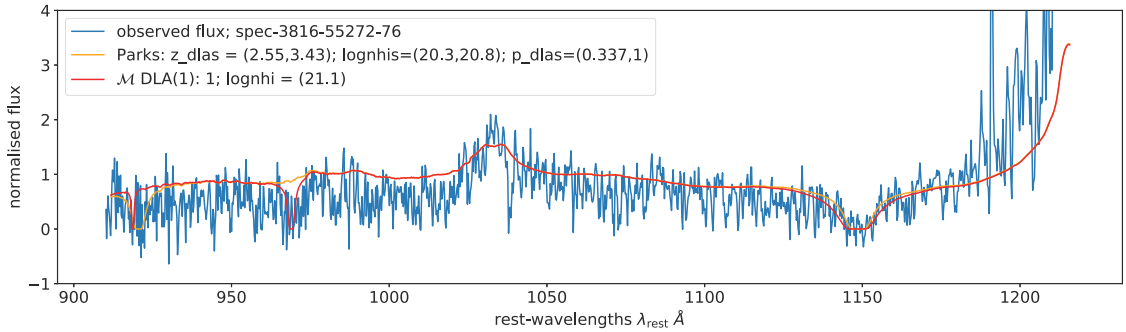

Figure 6. Blue: The normalized observed flux. The spectral ID represents `spec-plate-mjd-fiber_id`. Yellow: Parks' predictions on top of our null model. Our model predicts only one DLA while the CNN model in Parks et al. (2018) predicts two DLAs. One of the DLAs predicted by Parks et al. (2018) is coincident with the Ly γ absorption from our predicted DLA. `z_dla` corresponds to the DLA redshifts reported in Parks' catalogue, and `lognhi` corresponds to the column density estimations of Parks' catalogue. `p_dla` is the `dla_confidence` reported in Parks. Red: Our current model with the highest model posterior and the MAPs of column densities. In this spectrum, we show that it is crucial to include Lyβ and Lyγ absorption from the DLA in the DLA profile. It not only helps to localize the DLA, but it also predicts $N_{\text{H I}}$ more accurately using information from the Ly β region. The blue line shows the observed flux, the red curve is our multi-DLA GP prior, and the orange curve shows the predicted DLAs from Parks et al. (2018) subtracted from our mean model.

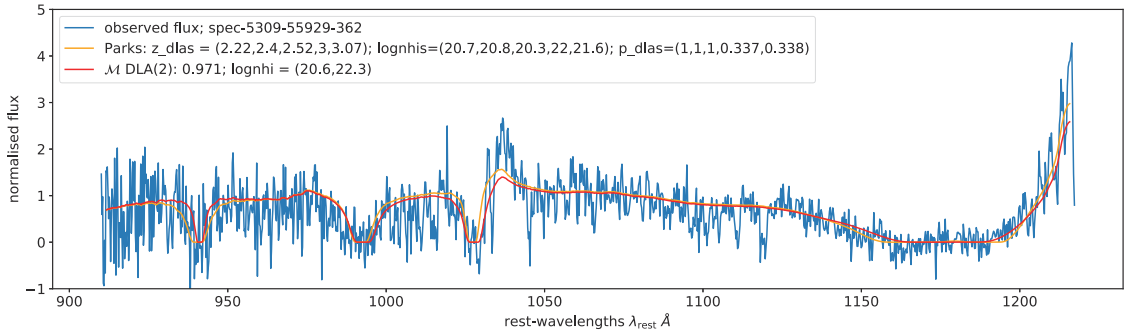

Figure 7. A spectrum in which we detect two DLAs. Blue: Normalized flux. Red: GP mean model with two intervening DLAs. Yellow: The predictions from Parks' catalogue. Pink: The MAP prediction of Garnett et al. (2017) on top of the GP mean model without mean flux suppression. The model posterior from Garnett et al. (2017) is listed in the legend (1) with the MAP value of $\log_{10} N_{\text{H I}}$. The column density estimate for the DLA near $\lambda_{\text{rest}} = 1\,025$ Å has large uncertainty (see Fig. 8). It is thus possible that this DLA could be a sub-DLA, as preferred by Parks et al. (2018).

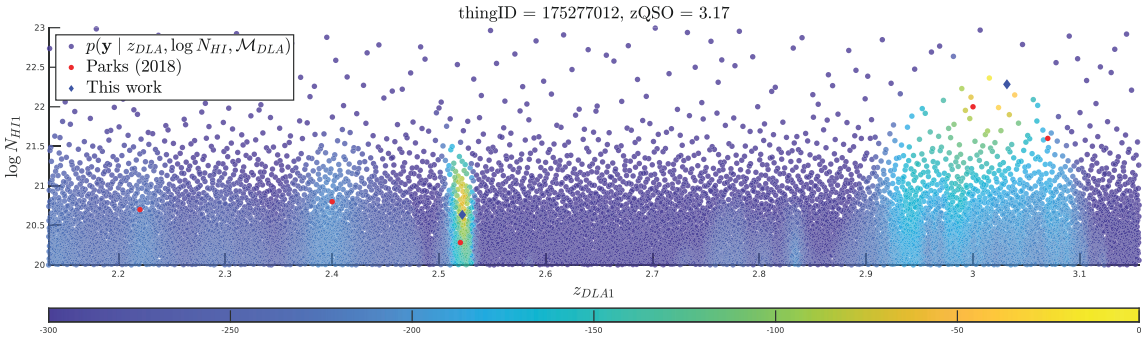

Figure 8. The log sample likelihoods for the DLA model of the spectrum shown in Fig. 7, normalized to range from $-\infty$ to 0. The DLA at $z_{\text{DLA}} \sim 2.52$ could be a sub-DLA (as preferred by Parks et al. 2018), as the $\log_{10} N_{\text{H I}}$ estimate is uncertain. However, we found that the two-DLA model posterior $\log p(\mathcal{M}_{\text{DLA}(2)} \mid \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}) = -638$ is still higher than the model posterior from combining one-DLA and one-sub-DLA, which is $\log p(\mathcal{M}_{\text{DLA}(1)} + \mathcal{M}_{\text{sub}} \mid \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}) = -691.47$.

dN/dX, and total column densities Ω_{DLA}) of the DLAs detected in our catalogue.

10.1 ROC analysis

To evaluate how well our multi-DLA classification reproduces earlier results, we rank our DLA detections using the log posterior odds between the DLA model (summing up all possible DLA models $\{\mathcal{M}_{\text{DLA}(i)}\}_{i=1}^{k}$) and the null model:

$$\log(\text{odds}) = \log \Pr(\{\mathcal{M}_{\text{DLA}}\} \mid \mathcal{D}, z_{\text{QSO}}) - \log \Pr(\mathcal{M}_{\neg\text{DLA}} \mid \mathcal{D}, z_{\text{QSO}}), \quad (66)$$

where the ranking is over all sightlines. From the top of the ranked list based on the log posterior odds, we calculate the true positive

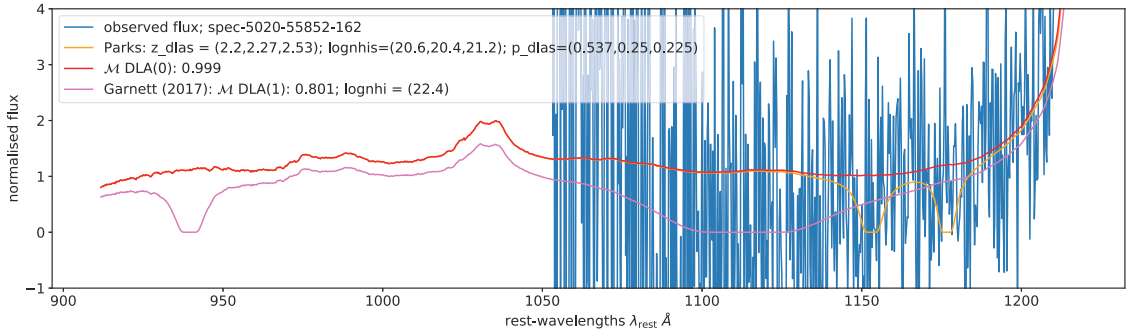

Figure 9. A noisy spectrum at $z_{\text{QSO}} = 2.378$ fitted with a large DLA by Garnett et al. (2017). Red: The model presented in this paper predicts no DLA detection in this spectrum. Pink: The MAP prediction of Garnett et al. (2017) on top the GP mean model without the mean-flux suppression. Gold: The prediction of Parks et al. (2018) subtracted from our mean model. Note that Parks et al. (2018) also indicate a detection of a DLA at $z_{\text{DLA}} = 2.53$, but outside the range of this spectrum.

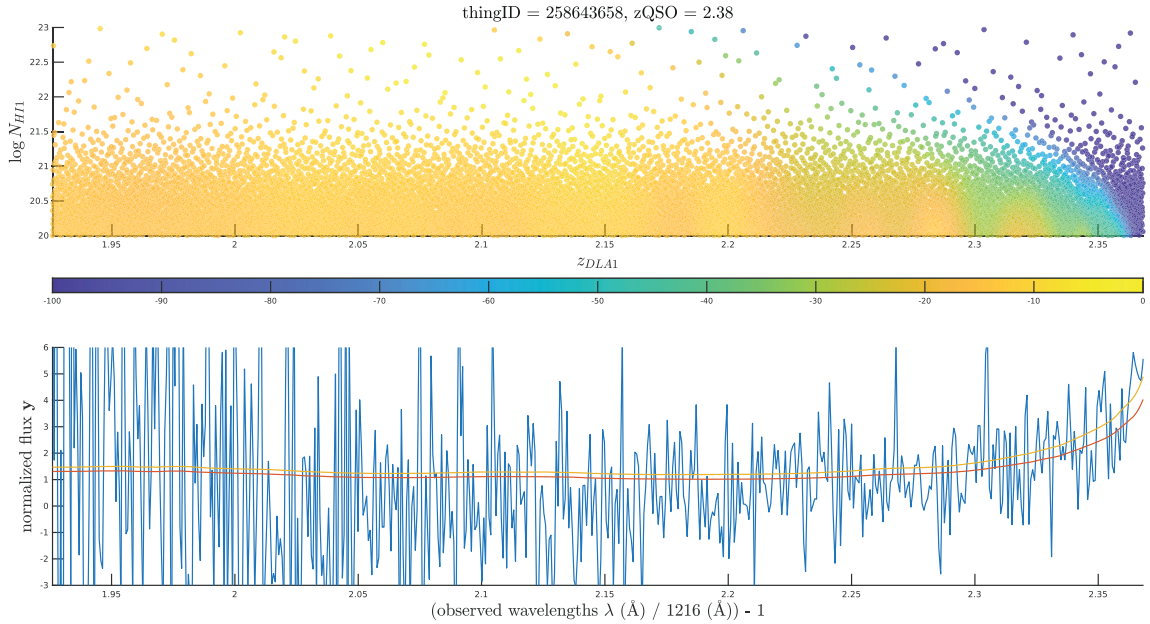

Figure 10. Top: The sample likelihoods of the spectrum shown in Fig. 9. The colour bar indicates the normalized log likelihoods ranging from $-\infty$ to 0. Bottom: The orange curve indicates the GP mean model before mean-flux suppression, the red curve represents the mean model after suppression, and the blue line is the normalized flux of this spectrum. The x-axis of this spectrum is rescaled to be the same as the z_{DLA} presented in the upper panel.

rate and false positive rate for each rank:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}};$$
$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \quad (67)$$

The true positive rate is the fraction of sightlines where we detect DLAs (ordered by their rank) divided by the number of sightlines with DLAs detected by earlier catalogues. The false positive rate is the number of detections of DLAs divided by the number of sightlines where earlier catalogues did not detect DLAs. In Fig. 11, we show the TPR and FPR in a receiver-operating characteristics (ROC) plot to show how well our classification performs. We have compared to the concordance DLA catalogue (Lee et al. 2013) in the hope that it approximates ground truth, there being no completely reliable DLA catalogue.

We also want to know how well our pipeline can identify the number of DLAs in each spectrum. The DR9 concordance catalogue does not count multiple DLA spectra, and so we compare our multi-DLA detections to the catalogue published by Parks et al. (2018). Each DLA detected in Parks et al. (2018) comes with a measurement of their confidence of detection (`dla_confidence` or $p_{\text{DLA}}^{\text{Parks}}$) and an MAP redshift and column density estimate. We compare our multiple DLA catalogue to those spectra with $p_{\text{DLA}}^{\text{Parks}} > 0.98$. The resulting ROC plot is shown in Fig. 12. We count a maximum of two DLAs in each spectrum: three or more DLAs in a single sightline are extremely rare and do not provide a large enough sample for an ROC plot. Parks' catalogue is not a priori more reliable than ours, especially in spectra with multiple DLAs, but comparing the first two DLAs is a reasonable way to validate our method's ability to detect multiple DLAs.

These spectra are counted by breaking down each two-DLA sightline (either in Parks or our catalogue) into two single observations. For example, if there are two DLAs detected in Parks and one DLA detected in our pipeline for an observation $\mathcal{D}$, we will assign one

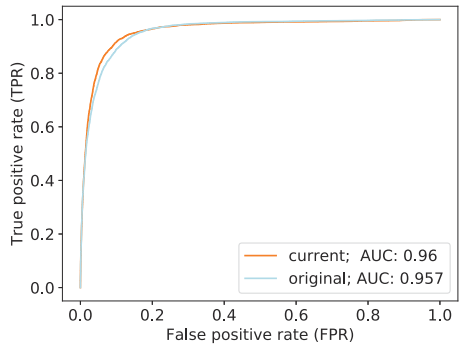

Figure 11. The ROC plot made by ranking the sightlines in BOSS DR9 samples using the log posterior odds of containing at least one DLA. Ground truths are from the DR9 concordance catalogue. The orange curve shows the ROC plot of our current multi-DLA model, and the blue curve is derived from Garnett et al. (2017). In this plot, we consider only the model containing at least one DLA $p(\{\mathcal{M}_{\rm DLA}\} \mid \mathcal{D})$, rather than the multiple DLA models, as the concordance catalogue contains only one DLA per spectrum.

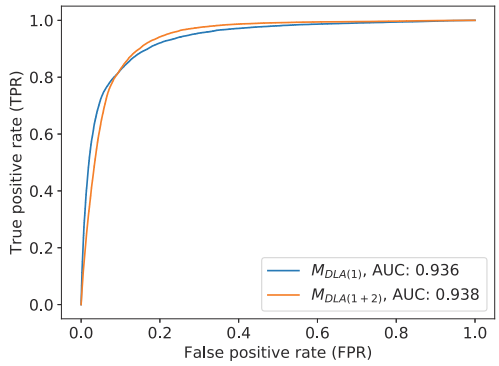

Figure 12. The ROC plot for sightlines with one and two DLA detections, by using the catalogue of Parks et al. (2018) (with `dla_confidence` >0.98) as ground truth.

ground-truth detection to $p(\mathcal{M}_{\rm DLA(1)} \mid \mathcal{D})$ and assign one ground-truth detection to $p(\mathcal{M}_{\neg \rm DLA} \mid \mathcal{D})$. On the other hand, if there is only one DLA detected in Parks and two DLAs detected in our pipeline, we will assign one ground-truth detection to $p(\mathcal{M}_{\rm DLA(2)} \mid \mathcal{D})$ and one ground-truth non-detection to $p(\mathcal{M}_{\rm DLA(2)} \mid \mathcal{D})$.

In Fig. 13, we also analyse the MAP estimate of the parameters $(z_{\rm DLA}, \log_{10} N_{\rm H\,I})$ by comparing with the reported values in DR9 concordance DLA catalogue. The median difference between these two is -2.2×10^{-4} (-66.6 km s^{-1}) and the interquartile range is 2.2×10^{-3} (662 km s^{-1}). For the log column density estimate, the median difference is 0.040, and the interquartile range is 0.26. The medians and interquartile ranges of the MAP estimate are very similar to the values reported in Garnett et al. (2017) with the median of $z_{\rm DLA}$ slightly smaller and the median of $\log_{10} N_{\rm H\,I}$ slightly larger. Note that the DR9 concordance catalogue is not the ground truth, so small variations in comparison to Garnett et al. (2017) can be considered to be negligible. As shown in Fig. 13, both histograms are roughly diagonal, although the scatter in column density MAP is large. Note that our DLA-detection procedure is designed to evaluate the model evidence across all of parameter space: a single sample MAP cannot convey the full posterior probability distribution. In Section 10.2, we thus describe a procedure to propagate the posterior density in the parameter space directly to column density statistics.

10.2 CDDF analysis

We follow Bird et al. (2017) in calculating the statistical properties of the modified DLA catalogue presented in this paper. We summarize the properties of DLAs using the averaged binned CDDF, the incident probability of DLAs (dN/dX), and the averaged matter density as a function of redshift ($\Omega_{\rm DLA}(z)$).

To plot these summary statistics, we need to convert the probabilistic detections in the catalogue to the expected average number of DLAs and their corresponding variances. We first describe how we compute the expected number of DLAs in a given column density and redshift bin. Next, we show how we derive the CDDF, dN/dX, and $\Omega_{\rm DLA}(z)$ from the expected number of DLAs. A sample of n observed spectra contains a sequence of n model posteriors $p^1_{\rm DLA}, p^2_{\rm DLA}, ..., p^n_{\rm DLA}$ defined by

$$p^i_{\rm DLA} = p(\{\mathcal{M}_{\rm DLA}\} \mid \boldsymbol{y}_i, \boldsymbol{\lambda}_i, \boldsymbol{\nu}_i, z_{{\rm QSO}i}), \tag{68}$$

where $i = 1, 2, ..., n$ is the index of the spectrum, and the DLA model here includes all computed DLA models $\{\mathcal{M}_{\rm DLA}\} = \{\mathcal{M}_{{\rm DLA}(i)}\}^k_{i=1}$, so that $k = 4$ is the maximum possible number of DLAs in each spectrum in our model.

Suppose the region of interest is in a specific bin $\boldsymbol{\Theta}$, an interval in the parameter space of column density or DLA redshift $\boldsymbol{\Theta} \in \{N_{\rm H\,I}, z_{\rm DLA}\}$. To compute the posterior of having DLAs in each spectrum in a given bin $\boldsymbol{\Theta}$, $p^i_{\rm DLA}(\{\mathcal{M}_{\rm DLA}\} \mid \boldsymbol{\Theta})$, we integrate over the sample likelihoods in the bin and multiply the model posterior by the total $p^i_{\rm DLA}$ for spectrum i:

$$p^i_{\rm DLA}(\{\mathcal{M}_{\rm DLA}\} \mid \boldsymbol{\Theta}) \propto p^i_{\rm DLA} \times \int_{\underline{\boldsymbol{\Theta}}}^{\overline{\boldsymbol{\Theta}}} p(\boldsymbol{y}_i \mid \{\mathcal{M}_{\rm DLA}\}, \boldsymbol{\lambda}_i, \boldsymbol{\nu}_i, z_{{\rm QSO}i}, \theta) {\rm d}\theta \,. \tag{69}$$

θ is either $z_{\rm DLA}$ or $\log_{10} N_{\rm H\,I}$ and $\theta \in \boldsymbol{\Theta} = (\underline{\boldsymbol{\Theta}}, \overline{\boldsymbol{\Theta}})$.

We calculate the posterior probability of having N DLAs by noting that the full likelihood follows the Poisson–Binomial distribution. Consider a sequence of trials with a probability of success equal to $p^i_{\rm DLA}(\{\mathcal{M}_{\rm DLA}\} \mid \boldsymbol{\Theta}) \in [0, 1]$. The probability of having N DLAs out of a total of n trials is the sum of all possible N DLAs subsets in the whole sample:

$$\Pr(N) = \sum_{{\rm DLA} \in F_N} \prod_{i \in {\rm DLA}} p^i_{\rm DLA}(\{\mathcal{M}_{\rm DLA}\} \mid \boldsymbol{\Theta}) \times \prod_{j \in {\rm DLA}^c} (1 - p^j_{\rm DLA}(\{\mathcal{M}_{\rm DLA}\} \mid \boldsymbol{\Theta})), \tag{70}$$

where F_N corresponds to all subsets of N integers that can be selected from the sequence $\{1, 2, ..., n\}$. The above expression means we select all possible N choices from the entire sample, calculate the probability of those N choices having DLAs and multiply that by the probability of the other $n - N$ choices having no DLAs. If all $p^i_{\rm DLA}(\{\mathcal{M}_{\rm DLA}\} \mid \boldsymbol{\Theta})$ are equal, the Poisson–Binomial distribution reduces to a Binomial distribution.

The above Poisson–Binomial distribution is not trivial to compute given our large sample size. The technical details of how to evaluate equation (70) efficiently are described in Bird et al. (2017). In short, we use Le Cam (1960)'s theorem to approximate those spectra with $p^i_{\rm DLA}(\{\mathcal{M}_{\rm DLA}\} \mid \boldsymbol{\Theta}) < p_{\rm switch} = 0.25$ by an ordinary Poisson distribution, and evaluate the remaining samples with the discrete Fourier transform (Fernandez & Williams 2010). Our catalogue contains the posteriors of samples in a given spectrum. Combined with the above probabilistic description of the total number of DLAs in the entire sample, we are able to obtain not only the point estimation of Pr(N) but also its probabilistic density interval.

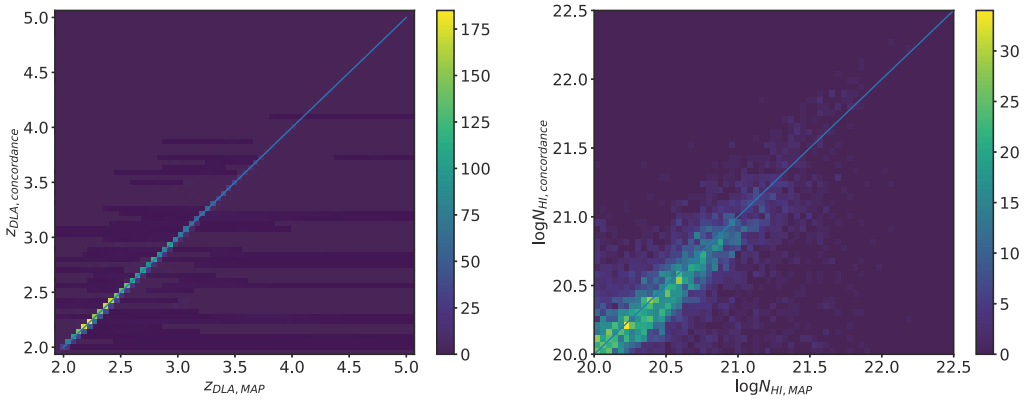

Figure 13. The MAP estimates of the DLA parameters $\theta = (z_{\rm DLA}, \log_{10} N_{\rm H\,I})$ for DLAs detected by our model in spectra observed by SDSS DR9, compared to the values reported in the concordance catalogue. The straight line indicates a perfect fit. Note that the concordance $\log_{10} N_{\rm H\,I}$ values are not ground truth, so the scatter in column density predictions was expected.

We thus compute the CDDF in a given bin $\mathbf{\Theta} = N_{\rm H\,I} \in [N_{\rm H\,I}, N_{\rm H\,I} + \Delta N_{\rm H\,I}]$ with

$$f(N) = \frac{F(N)}{\Delta N \Delta X(z)}, \tag{71}$$

where $F(N) = \mathbb{E}(N \mid N_{\rm H\,I} \in [N_{\rm H\,I}, N_{\rm H\,I} + \Delta N_{\rm H\,I}])$ is the expected number of absorbers at a given sightline within a column density interval. Thus, the CDDF $f(N)$ is the expected number of absorbers per unit column density per unit absorption distance, within a given column density bin.

The definition of absorption distance $\Delta X(z)$ is

$$X(z) = \int_0^z (1+z')^2 \frac{H_0}{H(z')} {\rm d}z', \tag{72}$$

which includes the contributions of the Hubble function $H^2(z)/H_0^2 = \Omega_{\rm M}(1+z)^3 + \Omega_\Lambda$, with $\Omega_{\rm M}$ the matter density and Ω_Λ the dark energy density.

The incident rate of DLAs ${\rm d}N/{\rm d}X$ is defined as

$$\frac{{\rm d}N}{{\rm d}X} = \int_{10^{20.3}}^{\infty} f(N \mid N_{\rm H\,I}, X \in [X, X + {\rm d}X]) {\rm d}N_{\rm H\,I}, \tag{73}$$

which is the expected number of DLAs per unit absorption distance.

The total column density $\Omega_{\rm DLA}$ is defined as

$$\Omega_{\rm DLA} = \frac{m_{\rm P} H_0}{c \rho_{\rm c}} \int_{10^{20.3}}^{\infty} N_{\rm H\,I} f(N \mid N_{\rm H\,I}, X \in [X, X + {\rm d}X]) {\rm d}N_{\rm H\,I}, \tag{74}$$

where $\rho_{\rm c}$ is the critical density at $z = 0$ and $m_{\rm P}$ is the proton mass.

10.3 Statistical properties of DLAs

Based on the above calculations, we show our CDDF in Fig. 14, $\frac{{\rm d}N}{{\rm d}X}$ in Fig. 15, and $\Omega_{\rm DLA}$ in Fig. 16.[7] Note that for determining the statistical properties of DLAs, we limit the samples of $z_{\rm DLA}$ to the range redward of the Lyman β in the QSO rest frame, as in Bird et al. (2017).

Fig. 14 shows the CDDF from our DR12 catalogue in comparison to the DR9 catalogue of Noterdaeme et al. (2012). Our CDDF analysis combines all spectral paths with QSO redshift smaller than 5, $z_{\rm DLA} < 5$. The CDDF statistics are dominated by the low-redshift

[7]The table files to reproduce Figs 14–16 will be posted in http://tiny.cc/multidla_catalog_gp_dr12q.

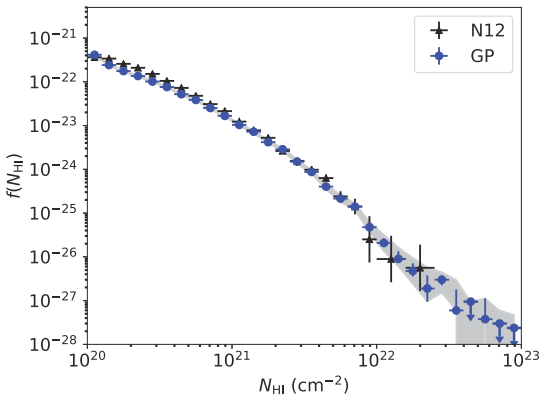

Figure 14. The CDDF based on the posterior densities for at least one DLA (blue, 'GP'). The DLAs are derived from SDSS DR12 spectra using the method presented in this paper. We integrate all spectral lengths with $z < 5$. We also plot the CDDF of Noterdaeme et al. (2012) (N12; black) as a comparison. The error bars represent the 68 per cent confidence limits, while the grey filled band represents the 95 per cent confidence limits. Note that our CDDF completely overlaps with those of N12 for column densities in the range 10^{21} cm^{-2} $< N_{\rm H\,I} < 10^{22}$ cm^{-2}.

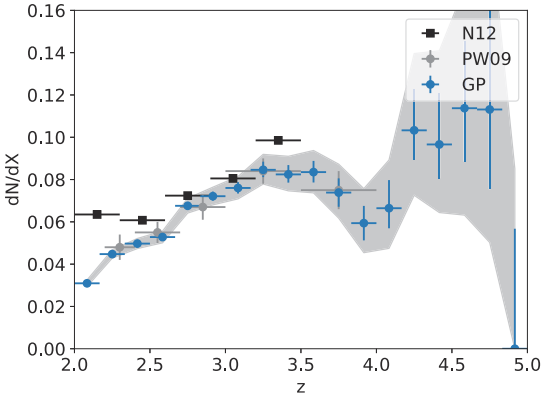

Figure 15. The line density of DLAs as a function of redshift from our DR12 multi-DLA catalogue (blue, 'GP'). We also plot the results of Noterdaeme et al. (2012) (N12; black) and Prochaska & Wolfe (2009) (PW09; grey). Note that statistical error was not computed in Noterdaeme et al. (2012).

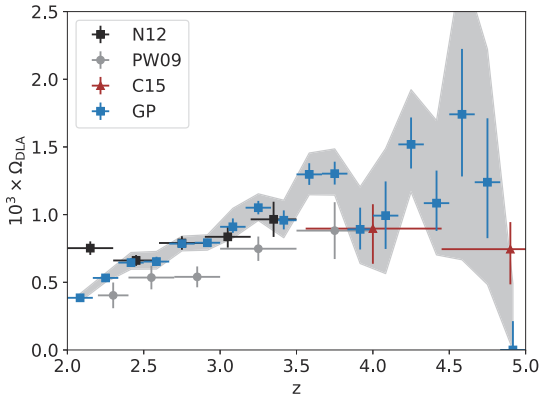

Figure 16. The total H I density in DLAs, $\Omega_{\rm DLA}$, from our DR12 multi-DLA catalogue as a function of redshift (blue, 'GP'), compared to the results of Noterdaeme et al. (2012) (N12; black), Prochaska & Wolfe (2009) (PW09; grey), and Crighton et al. (2015) (C15; red).

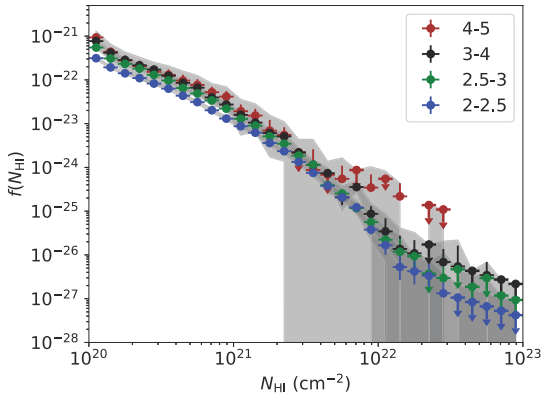

Figure 17. The redshift evolution (or non-evolution) of the CDDF. Labels show the absorber redshift ranges used to plot the CDDFs. In column density and redshift ranges with no detection at 68 per cent confidence, a down-pointing arrow is shown indicating the 68 per cent upper limit.

absorbers, as demonstrated in Fig. 17. The error bars represent the 68 per cent confidence interval, while the grey shaded area encloses the 95 per cent highest density region. The CDDF values in Fig. 14 are calculated from the posterior distribution directly. We note that there are only two DLAs with MAP $\log_{10}N_{\rm H\,I} > 22.5$ in our catalogue with high confidence ($p_{\rm DLA} > 0.99$). The non-zero values in the CDDF are due to uncertainty in $\log_{10}N_{\rm H\,I}$, not positive detections.

Noterdaeme et al. (2012) contains multi-DLAs, but, as described in section 2.2 in their paper, they applied a stringent cut on their samples with CNR >3, where CNR refers to the continuum-to-noise ratio. The CDDF of N12 in Fig. 14 is thus a subsample of their catalogue. We, on the other hand, use all data even those with low signal-to-noise ratios. Comparing to our previously published CDDF (Bird et al. 2017), the CDDF in this paper shows DLA detections at low $N_{\rm H\,I}$ are consistent with Noterdaeme et al. (2012). Introducing the sub-DLA as an alternative model successfully regularizes detections at $\sim 10^{20}$ cm^{-2}.[8]

Fig. 15 shows the line density of DLAs. Our results are again consistent with those of Prochaska & Wolfe (2009) and Noterdaeme et al. (2012) where they both agree. Our detections are between those two catalogues at low-redshift bins and consistent with Prochaska & Wolfe (2009) in the highest redshift bin. Comparing to our previous dN/dX (Bird et al. 2017), we moderately regularize the detections of DLAs at high redshifts. This change shows that changing the mean model of the GP to include the mean flux absorption prevents the pipeline confusing the suppression due to the Lyman alpha forest with a DLA. While the change of posterior modes in dN/dX is large at high-redshift bins, we note that those changes are mostly within 95 per cent confidence interval of our previously published line densities. All analyses shown measure a peak in dN/dX at $z \sim 3.5$. This may be partially due to $z_{\rm DLA} = 3.5$ the SDSS colour selection algorithm systematic identified by Prochaska, Worseck & O'Meara (2009), which oversamples Lyman-limit systems (LLS), especially near the QSO, in the redshift range 3.0–3.6 (Worseck & Prochaska 2011; Fumagalli et al. 2013). Note, however, that in our analysis neighbouring redshift bins are highly correlated and so a statistical fluctuation is also a valid explanation. We have checked visually that our sub-DLA model successfully models spectra with an LLS in the proximate zone of the QSO emission peak.

Fig. 16 shows the total column density $\Omega_{\rm DLA}$ in DLAs in units of the cosmic density. Our results are mostly consistent with Noterdaeme et al. (2012) although we have slightly lower $\Omega_{\rm DLA}$ at $z \sim 2$. This is due to our Occam's razor penalty, which suppresses DLAs in spectra which are not long enough to include the full width of the DLA. Since these are all low-redshift QSOs, this suppresses DLA detections at $z < 2.3$. As discussed in Noterdaeme et al. (2012), Sánchez-Ramírez et al. (2016), and Bird et al. (2017), the relatively low $\Omega_{\rm DLA}$ of Prochaska & Wolfe (2009) is due to the smaller sample size of the SDSS DR5 data set. We also compare our $\Omega_{\rm DLA}$ to that measured by Crighton et al. (2015) at high redshifts ($z = 4$ and $z = 5$). Crighton et al. (2015) used a small but higher signal-to-noise data set. Our results at $z = 4$ and $z = 5$ are consistent with those from Crighton et al. (2015). However, we note that the relatively small sample of Crighton et al. (2015) may bias it slightly low, as contributions from DLAs with $N_{\rm H\,I}$ higher than expected to be in the survey will not be included in their $\Omega_{\rm DLA}$ estimate. Our Bayesian analysis includes possible contributions of undetected DLAs with column density up to $\log_{10}N_{\rm H\,I} = 23$ in the error bars via the prior on the column density.

Compared to our previously published $\Omega_{\rm DLA}$ (Bird et al. 2017), we found a reduction in $\Omega_{\rm DLA}$ between $z = 4$ and $z = 5$. This is due to the incorporation of a better mean flux vector model, which reduces the posterior density of high column density systems for high-redshift absorbers (although within the 95 per cent confidence bars of the earlier work). Our confidence intervals are also substantially smaller for $z_{\rm DLA} \gtrsim 3.7$ than in (Bird et al. 2017). This is due to our inclusion, for the first time, of information from the Lyman β absorption of the DLAs, which both constrains DLA properties and helps to distinguish DLAs from noise fluctuations.

We have tested the robustness of our method with respect to spectra with different SNRs and found that, as in Bird et al. (2017), the statistical properties predicted by our method are uncorrelated with the QSO SNR. Furthermore, the presence of a DLA is uncorrelated with the QSO redshift, fixing a statistical systematic in the earlier work.

As a cross-check of our wider catalogue, we also tested the CDDF, line densities, and total column densities of the DLAs in our catalogue with a full range of $z_{\rm DLA}$, from Ly ∞ to Ly α. The CDDF was very similar to the CDDF excluding the Ly β region shown in Fig. 14, but with a moderate increase at high column density. dN/dX was almost

[8]Note again the artefact at $\sim 10^{20}$ cm^{-2} will not affect the analyses of dN/dX or $\Omega_{\rm DLA}$ as the definition of a DLA is absorbers with $N_{\rm H\,I} > 10^{20.3}$ cm^{-2}.

identical to Fig. 15, indicating that the detection of DLAs is robust even though we extend our sampling range to Ly ∞. However, $\Omega_{\rm DLA}$ increases for $3.5 < z_{\rm DLA} < 4.0$. By visual inspection we found that this is due to the spectra where the QSO redshift from the SDSS pipeline in error and a Lyman break trough appears at the blue end of the spectrum in a region the code expects to contain only Ly β absorption. As our model does not account for redshift errors, it explains the absorption due to these troughs by DLAs.

10.4 Comparison to Garnett's catalogue

To understand the effect of the modifications we made to our model in this paper, we visually inspected a subset of spectra with high model posteriors of a DLA in Garnett et al. (2017) ($p_{\rm DLA}^{\rm Garnett}$) but low model posteriors in our current model ($p_{\rm DLA}$). In particular, we chose spectra with ($p_{\rm DLA}^{\rm Garnett} - p_{\rm DLA} > 0.99$).

A large fraction of these spectra falls within the Ly β emission region. One plausible explanation is that the Ly β emission region has a higher noise variance, which makes it harder to distinguish the DLA and sub-DLA models. We also checked that we are not unfairly preferring the sub-DLA model during model selection. Our model selection uses the sub-DLA model only to regularize the DLA model and does not consider cases where DLAs and sub-DLAs occur in the same spectrum. Thus a spectrum with a clear detection of a sub-DLA could fail to detect a true DLA at a different redshift. In light of this, we also tested if combining multi-DLA models with a sub-DLA affects our results.

We modified the DLA model, assuming that the DLA and sub-DLA models are independent, to include the sub-DLA model prior. We then considered an iterative sampling procedure: First, we sampled the k-th DLA likelihood. Next we used the k-th DLA parameter posterior as a prior to sample $\mathcal{M}_{\rm DLA(k)}$ and combine $\mathcal{M}_{\rm DLA(k)}$ with the sub-DLA model via sampling a non-informative prior. The full procedure can be written as

$$p(\{\theta_i\}_{i=1}^{k} \mid \mathcal{M}'_{\rm DLA(k)}, \mathcal{D}, z_{\rm QSO}) = (1 + p(\theta_{\rm sub} \mid \mathcal{M}_{\rm sub}, z_{\rm QSO})) \times p(\{\theta_i\}_{i=1}^{k} \mid \mathcal{M}_{\rm DLA(k)}, \mathcal{D}, z_{\rm QSO}), \quad (75)$$

For computational simplicity, we only consider the modified model until $\mathcal{M}'_{\rm DLA(3)}$; the probability of $\mathcal{M}'_{\rm DLA(4)}$ is expected to be insignificant comparing to the total DLA model posterior, $p(\{\mathcal{M}_{\rm DLA}\} \mid \mathcal{D}, z_{\rm QSO})$.

In practice, however, we found that this made a small difference to our results, only marginally modifying the ROC curve and CDDF. Moreover, the ability of the sub-DLA model to regularize low column density DLAs was reduced, so we have preserved our default model.

10.5 Comparison to Parks catalogue

In this section, we compare our results with Parks et al. (2018). We first show the differences between our MAP predictions and Parks' predictions for DLA redshift and column density. We required $p_{\rm DLA}^{\rm Parks} > 0.98$. We measured the difference in posterior parameters when both pipelines predicted one DLA. As shown in Fig. 18, both histograms are roughly symmetric. We measure small median offsets between two pipelines with

$$\begin{aligned} \text{median}(z_{\rm DLA}^{\rm MAP} - z_{\rm DLA}^{\rm Parks}) &= 0.00010; \\ \text{median}(\log N_{\rm H\,I}^{\rm MAP} - \log N_{\rm H\,I}^{\rm Parks}) &= 0.016. \end{aligned} \quad (76)$$

We also compared our absorber redshift measurements and column density measurements to Parks' catalogue for those spectra that we both agree contain two DLAs. The differences between these two have small median offsets of $\Delta z_{\rm DLA} = 0.000052$ and $\Delta \log_{10} N_{\rm H\,I} = 0.006$ (and dominated by low column density systems).

We show the disagreement between multi-DLA predictions for our catalogue and Parks' catalogue in Table 1. Note that though the multi-DLA detections between our method and Parks do not completely agree, the level of disagreement is small: 6.1 per cent. Moreover, if Parks predicts one or two DLAs, our method generally detects one or two DLAs. There are, however, some spectra where we detected >2 DLAs, but Parks detected none. To understand the statistical effect of this discrepancy, we compare our DLA properties to those reported by Parks et al. (2018). We plot the CDDF and dN/dX of that catalogue. We assume $p_{\rm DLA}^{\rm Parks} > 0.9$ represents a DLA and use $z_{\rm DLA}$ and $\log_{10} N_{\rm H\,I}$ reported in their catalogue in JSON format.[9] To compute the sightline path searched over, we assume their CNN model was searching the range Ly ∞ to Ly α in the QSO rest frame. Note this differs slightly from Parks et al. (2018), section 3.2, where a sightline search radius ranging from 900 to 1346 Å in the QSO rest frame is given. However, we know the centres of DLAs should be at a redshift between Ly ∞ and Ly α in the rest frame and modify our search paths accordingly.

Fig. 19 shows that dN/dX is consistent with Noterdaeme et al. (2012) for $z_{\rm DLA} < 3.5$ (although lower than our measurement at higher redshift). The CNN is thus successfully detecting DLAs, especially the most common case of DLAs with a low column density. There are fewer DLAs detected at higher redshift, likely reflecting the increased difficulty for the CNN of distinguishing DLAs from the Lyman α forest. This is discussed in Parks et al. (2018), who note that the CNN finds it difficult to detect a weak DLA in noisy spectra. However, as shown in Fig. 20, the CDDF measured by the CNN model is significantly discrepant with other surveys for large column densities. Note that the scale is logarithmic: the CNN is failing to detect >60 per cent of DLAs with $\log_{10} N_{\rm H\,I} > 21$. We noticed that large DLAs were often split into two objects with lower column density, which accounts for many of the discrepancies between our two data sets. We suspect this might be due to the limited size of the convolutional filters used by Parks et al. (2018). If the filter is not large enough to contain the full damping wings of a given DLA, the allowed column density would be artificially limited.

11 CONCLUSION

We have presented a revised pipeline for detecting DLAs in SDSS QSO spectra based on Garnett et al. (2017). We have extended the pipeline to reliably detect up to four DLAs per spectrum. We have performed modifications to our model for the Lyman α forest to improve the reliability of DLA detections at high redshift and introduced a model for sub-DLAs to improve our measurement of low column density DLAs. Finally, we introduced a penalty on the DLA model based on Occam's razor which meant that spectra for which both models are a poor fit generally prefer the no-DLA model.

Our results include a public DLA catalogue, with several examples shown above and further examples easily plotted using a PYTHON package. We have visually inspected several extreme cases to validate our results and compared extensively to several earlier DLA catalogues: the DR9 concordance catalogue (Lee et al. 2013) and a DR12 catalogue using a CNN (Parks et al. 2018). Our new pipeline had very good performance validated against both catalogues.

[9] https://tinyurl.com/cnn-dlas

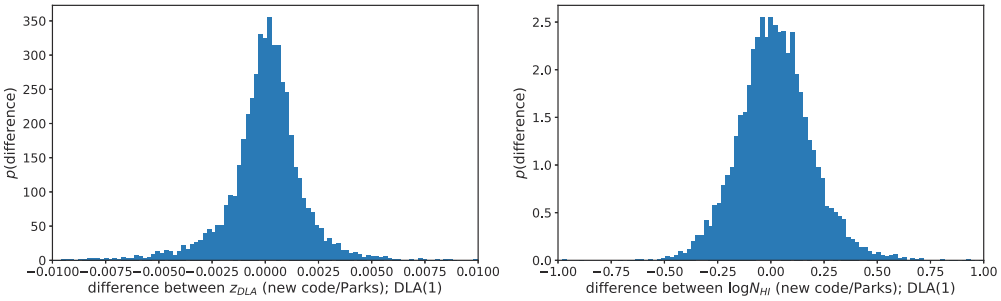

Figure 18. The difference between the MAP estimates of the DLA parameters $\theta = (z_{\rm DLA}, \log_{10} N_{\rm H\,I})$, against the predictions of Parks et al. (2018). We consider spectra that both catalogues agree contain one DLA.

Table 1. The confusion matrix for multi-DLAs detections between Garnett with multi-DLAs and Parks. Note we require both the model posteriors in Garnett and DLA confidence in Parks to be larger than 0.98. We also require $\log_{10} N_{\rm H\,I} > 20.3$.

Parks Garnett with multi-DLAs	0 DLA	1 DLA	2 DLAs	3 DLAs	4 DLAs
0 DLA	138 726	6197	142	6	0
1 DLA	3050	8752	335	4	0
2 DLAs	293	570	566	28	0
3 DLAs	30	39	34	21	0
4 DLAs	5	9	6	1	0

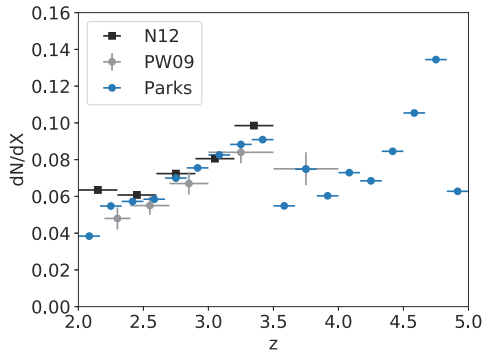

Figure 19. dN/dX from Parks et al. (2018). The dN/dX agrees well with other surveys, but there is a moderate deficit of DLAs at high redshifts.

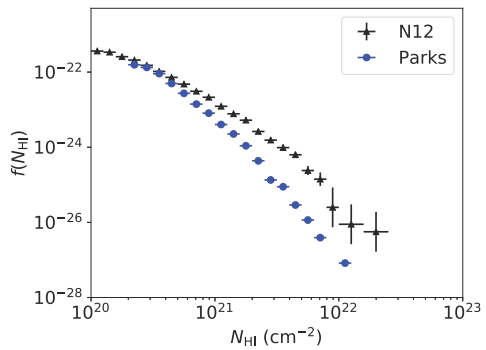

Figure 20. The CDDF from Parks et al. (2018), showing that the CNN algorithm substantially underestimates the number of DLAs in the high-$N_{\rm H\,I}$ regime.

Based on the revised pipeline, we also presented a new measurement of the abundance of neutral hydrogen from $z = 2$ to $z = 5$ using similar calculations to Bird et al. (2017). The statistical properties of DLAs were in good agreement with our previous results (Bird et al. 2017) and consistent with Noterdaeme et al. (2012), Prochaska & Wolfe (2009), and Crighton et al. (2015). The modifications made, including introducing a sub-DLA model, adjusting the mean flux, and penalizing complex models with Occam's razor, remove overdetections of low column density absorbers and make more robust predictions for the properties of DLAs at $z > 4$. Similarly to previous work, we detect only a small increase in the CDDF for $2 < z < 4$, and a similarly moderate increase in the line density of DLAs and $\Omega_{\rm DLA}$ over this redshift range.

ACKNOWLEDGEMENTS

The authors thank Bryan Scott and Reza Monadi for valuable comments. This work was partially supported by an AWS Machine Learning Research Awards allocation on EC2 and UCR's HPCC. SB was supported by the National Science Foundation (NSF) under award number AST-1817256. RG was supported by the NSF under award number IIS-1845434.

DATA AVAILABILITY

All the code to reproduce the data products is available in our GitHub repo: https://github.com/rmgarnett/gp_dla_detection/tree/master/multi_dlas. The final data products are available in this Google Drive: http://tiny.cc/multidla_catalog_gp_dr12q, including a MAT (HDF5) catalogue and a JSON catalogue. README files are included in each folder to explain the content of the catalogues.

REFERENCES

Becker G. D., Hewett P. C., Worseck G., Prochaska J. X., 2013, MNRAS, 430, 2067

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
Bird S., Garnett R., Ho S., 2017, MNRAS, 466, 2111
Carithers W., 2012, Published internally to SDSS
Cen R., 2012, ApJ, 748, 121
Crighton N. H. M. et al., 2015, MNRAS, 452, 217
Fauber L., Ho M.-F., Bird S., Shelton C. R., Garnett R., Korde I., 2020, MNRAS, preprint (arXiv:2006.07343)
Fernandez M., Williams S., 2010, IEEE Trans. Aerosp. Electron. Syst., 46, 803
Fumagalli M., O'Meara J. M., Prochaska J. X., Worseck G., 2013, ApJ, 775, 78
Fumagalli M., O'Meara J. M., Prochaska J. X., Rafelski M., Kanekar N., 2014, MNRAS, 446, 3178
Gardner J. P., Katz N., Weinberg D. H., Hernquist L., 1997, ApJ, 486, 42
Garnett R., Ho S., Bird S., Schneider J., 2017, MNRAS, 472, 1850
Haehnelt M. G., Steinmetz M., Rauch M., 1998, ApJ, 495, 647
Kim T.-S., Bolton J. S., Viel M., Haehnelt M. G., Carswell R. F., 2007, MNRAS, 382, 1657
Lee K.-G. et al., 2013, AJ, 145, 69
Le Cam L., 1960, Pac. J. Math., 10, 1181
Noterdaeme p. et al., 2012, A&A, 547, L1
Pâris I. et al., 2012, A&A, 548, A66
Pâris I. et al., 2014, A&A, 563, A54
Pâris I. et al., 2018, A&A, 613, A51
Parks D., Prochaska J. X., Dong S., Cai Z., 2018, MNRAS, 476, 1151
Pontzen A. et al., 2008, MNRAS, 390, 1349
Prochaska J. X., Wolfe A. M., 1997, ApJ, 487, 73
Prochaska J. X., Wolfe A. M., 2009, ApJ, 696, 1543
Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, ApJ, 635, 123
Prochaska J. X., Worseck G., O'Meara J. M., 2009, ApJ, 705, L113
Rasmussen C. E., Williams C. K. I., 2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, Cambridge, Massachusetts
Sánchez-Ramírez R. et al., 2016, MNRAS, 456, 4488
Slosar A. et al., 2011, J. Cosmol. Astropart. Phys., 2011, 001
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
Worseck G., Prochaska J. X., 2011, ApJ, 728, 23
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

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