Peering towards Cosmic Dawn
Proceedings IAU Symposium No. 333, 2018
V. Jelić & T. van der Hulst, eds.

doi:10.1017/S1743921317011322

Constraining Lyman continuum escape using Machine Learning

Sambit K. Giri[1*], Erik Zackrisson[2], Christian Binggeli[2], Kristiaan Pelckmans[3], Rubén Cubo[3] and Garrelt Mellema[1]

[1]Department of Astronomy and Oskar Klein Centre, Stockholm University, AlbaNova, SE-106 91 Stockholm, Sweden
[2]Department of Physics and Astronomy, Uppsala University, Box 516, SE-751 20 Uppsala, Sweden
[3]Department of Information Technology, Division of Systems and Control (Syscon), Uppsala University, Box 337, SE-751 05 Uppsala, Sweden
[*]email: sambit.giri@astro.su.se

Abstract. The James Webb Space Telescope (JWST) will observe the rest-frame ultraviolet/optical spectra of galaxies from the epoch of reionization (EoR) in unprecedented detail. While escaping into the intergalactic medium, hydrogen-ionizing (Lyman continuum; LyC) photons from the galaxies will contribute to the bluer end of the UV slope and make nebular emission lines less prominent. We present a method to constrain leakage of the LyC photons using the spectra of high redshift ($z \gtrsim 6$) galaxies. We simulate JWST/NIRSpec observations of galaxies at z =6–9 by matching the fluxes of galaxies observed in the Frontier Fields observations of galaxy cluster MACS-J0416. Our method predicts the escape fraction f_{esc} with a mean absolute error $\Delta f_{esc} \approx 0.14$. The method also predicts the redshifts of the galaxies with an error $\left\langle \frac{\Delta z}{(1+z)} \right\rangle \approx 0.0003$.

Keywords. Galaxies: high-redshift – dark ages, reionization, first stars – gravitational lensing: strong

1. Introduction

Models of galaxy-driven reionization require a relatively high average escape fraction of LyC photons ($f_{\rm esc} \approx 0.01-0.2$) in order to ionize the intergalactic medium (IGM) during the EoR (Mitra *et al.* 2015; Hartley and Ricotti 2016). While current measurements from local galaxies typically indicate $f_{\rm esc} < 0.1$ (e.g. Izotov *et al.* 2016), there have been a handful of objects observed with a high $f_{\rm esc}$ (> 0.4) at $z \approx 2-3$ (e.g. Vanzella *et al.* 2016). However, direct detections of LyC from the EoR are not possible due to the opacity of the neutral IGM. Various indirect methods have been proposed to measure f_{esc} at z> 6 using observations of gravitationally lensed objects (Jones *et al.* 2013; Zackrisson *et al.* 2013; Leethochawalit *et al.* 2016).

Zackrisson *et al.* (2013, 2017) propose to calculate the LyC escape fraction using spectroscopic data from the upcoming JWST. They argue that LyC photons that escape from galaxies will be unable to contribute to the photo-ionization of the interstellar medium. Therefore, galaxies with non-zero LyC escape fractions will display less prominent emission lines and bluer UV slopes. We present our simulations of observations of the JWST/NIRSpec instrument for galaxies similar to the ones seen in the lensed field of Frontier Field (FF) clusters†. We then use a machine learning regression algorithm

† www.frontierfields.org

for constraining the f_{esc} from these spectra. We extend the work done in Jensen *et al.* (2016) to consider the case where the redshift values of the galaxies are unknown.

2. Mock galaxy spectra

The FF program was initiated to observe faint high-redshift galaxies by combining the observations from Hubble Space Telescope (HST) and Spitzer Space Telescope (Lotz *et al.* 2017). These observations used six highly magnifying galaxy clusters which are MACS-J0416 (hereafter M0416), MACS-J0717, MACS-J1149, Abell-2744, Abell-S1063 and Abell-370. These galaxies will be targets for the JWST. We simulate spectra for the galaxies observed in the field of M0416 as this field is chosen in a proposal † for JWST Cycle 1 Guaranteed Time Observations (GTO) to be observed with the NIRSpec instrument. When this data will be available, we can use it to validate our study.

We use the CROC simulations (Gnedin and Kaurov 2014) to generate the star formation rates and the metallicity distribution of galaxies with stellar masses $M_\star \geqslant 10^7 M_\odot$. Approximately 1300 galaxies are simulated at redshift $z \approx 7$, 8, 9 and 10. Synthetic spectra are generated by using the Yggdrasil spectral synthesis code (Zackrisson *et al.* 2011). The nebular emissions were calculated by using the Cloudy code (Ferland *et al.* 2013). Finally, the effect on LyC is modelled using the description in Zackrisson *et al.* (2013) to obtain the spectra for various f_{esc}, for more detail see Zackrisson *et al.* (2017).

Our machine learning method requires a training set of spectra from different z and f_{esc}. To provide this, a redshift resolution of $\Delta z \approx 0.01$ is used to sample the redshifts 6 to 9. At each of these redshifts of a training set with spectra for $f_{esc} = 0.0, 0.1, ..., 1.0$ is produced. The closest spectra is chosen for each training set redshift from the dataset of synthetic spectra and the fluxes are scaled to match the magnitude observed in the FF catalogue ($m_{1500,AB} \approx 24 - 30$). Finally, the spectra are re-sampled to the JWST/NIRSpec resolution for $R = 100$ mode‡. Unfortunately, the currently planned GTO time is too short to allow for meaningful results given our current analysis. In the future, we will strive to improve our method to instead work with shorter exposure times. In the current study, we add noise corresponding to 10 hour exposures to our spectra using the recipe given in Jensen *et al.* (2016). Fig. 1 shows the normalized flux of the spectra with varying f_{esc}. We see a clear correlation between the escape fraction and certain spectral features, such as the UV slope and the emission lines.

3. Method

3.1. *Unshifting*

As we have spectra from different z, each nebular emission line will be observed at a different wavelength. Jensen *et al.* (2016) showed that the prediction of f_{esc} depends on the fluxes of these emission lines. In order to calculate f_{esc} we shift all the spectra to the rest frame wavelength.

We have a dictionary of *Gold Standard Spectra* (GSS), which are noise free spectra at the best possible wavelength resolution in the rest frame. The dictionary contains GSS with and without Lyman-α (Ly-α) emission line and with varying f_{esc}. Fluxes at wavelengths less than that of Ly-α line will be absorbed due to Gunn-Peterson effect (Gunn and Peterson 1965). When the column density of neutral hydrogen is high, part of the Ly-α flux are affected by the damping wing of the Gunn-Peterson absorption.

† CANUCS: www.stsci.edu/jwst/phase2-public/1208.pdf

‡ See https://jwst.stsci.edu/instrumentation/nirspec for the details of the observation modes.

https://doi.org/10.1017/S1743921317011322 Published online by Cambridge University Press

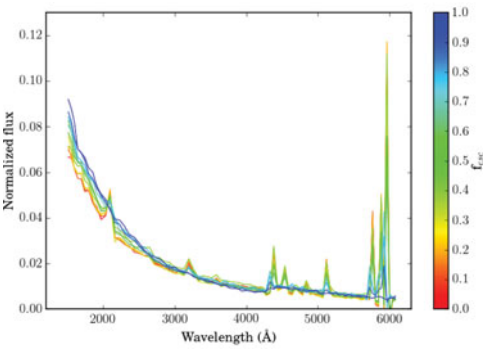

Figure 1. Synthetic spectrum of the simulated galaxies with varying escape fraction f_{esc}. Each spectra is normalized to have an area under the spectra equal to unity in the log-space. The horizontal axis gives the wavelength in the rest frame. Certain spectral features, such as the EW(Hβ) and UV slope, are seen to correlate with f_{esc}.

We shift all the observed spectra to the rest frame wavelength using a process we call *unshifting*. First, the wavelength resolution is matched by using nearest-neighbour interpolation. The observed spectra are shifted back to a GSS by moving the observed spectrum until the cross-correlation is maximum. This shifting process is repeated for all the GSS in the dictionary because the calculated *shift* is more accurate if the observed spectra shifts to the GSS that has similar features. For example, an observed spectra with no Ly-α emission line should shift to a GSS which does not have one.

3.2. *LASSO Regression*

The Least Absolute Shrinkage and Selection Operator (LASSO) is a machine learning method that fits a linear model to a dataset (Tibshirani 1996). The model fitted by LASSO is given by:

$$\hat{y} = \beta_0 + \sum_{i=1}^{N} \beta_i x_i \ , \tag{3.1}$$

where the m is the number of data points in the training set, which are $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_1, y_1)$,... $(\mathbf{x}_m, y_m)$. $\mathbf{x}_i$'s are the input vectors that have N features. Our features include fluxes from (N-1) wavelength bins and the *shift* value. y_i's are the labels which in this case will be either z or f_{esc}.

The value for the β's are determined by minimizing the least-square error $\sum_{j=1}^{m}[\hat{y}_j - y_j]^2$ along with a regularization term to control the over-fitting. LASSO uses an $\mathbf{L_1}$ regularization $\lambda \sum_{i=1}^{N} |\beta_i|$. λ is a tunable parameter called the *penalty*. We use cross-validation to estimate the *penalty*. The advantage of using $\mathbf{L_1}$ regularization is the imposition of sparsity into the model. We can see in Fig. 1 that there are few special features which are better correlated to the f_{esc} and for a good value of λ LASSO only gives non-zero coefficients to them. The λ can be determined using cross-validation (Kohavi *et al.* 1995). We have used 10-fold cross-validation throughout this study. We refer the interested reader to Ivezić *et al.* (2014) and Jensen *et al.* (2016) for more details.

4. Results

As we do not have the JWST observations now, we randomly select 20% of the $\approx$ 6800 training set galaxies and keep it as test set ($\approx$ 1700 galaxies). We use normalised fluxes of the spectra and the *shift* calculated during the *unshifting* as our input vectors. The

https://doi.org/10.1017/S1743921317011322 Published online by Cambridge University Press

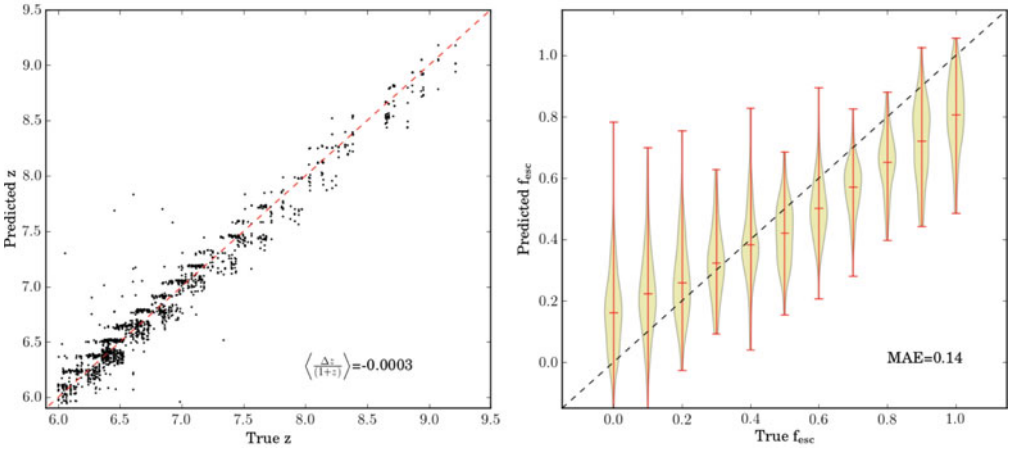

Figure 2. The prediction for z (left panel) and f_{esc} (right panel) is shown. *Left Panel*: The *unshifting* works well as the scatter is very close to the $z^{true} = z^{pred}$ line. *Right Panel*: The violin plot shows the prediction of f_{esc}. The markers on the violins are put at the mean and the two extrema. The width of each violin shows the probability density of predicting a value. The MAE for predicted f_{esc} is 0.14.

coefficient of the *shift* that LASSO predicts is particularly high, indicating that redshift is highly dependant on the *shift* calculation. Fig. 2 shows the performance of the fitted model on the test set. The left panel of Fig. 2 shows that the *unshifting* followed by a model fit by LASSO gives a good prediction for the z of the galaxies. The prediction error $\left\langle \frac{\Delta z}{(1+z)} \right\rangle$ is comparable to the value given in Castellano *et al.* (2016).

The nebular emission lines have a strong correlation with the f_{esc} whereas the coefficient for the *shift* value is zero as the f_{esc} is independent of it. The right panel in Fig 2 shows the prediction of f_{esc} for the test set. The width of the violin at each predicted f_{esc} represents the probability density for the value predicted for that true f_{esc}. All the plots show a unimodal distribution. This suggests that the model has a higher chance of predicting the mean (shown with the marker). The predictions have a larger spread in the lower true f_{esc}. However, the probability of predicting the outliers are quite low. LASSO fits a model for f_{esc} that is similar to the ones shown in Jensen *et al.* (2016). The mean absolute error is in agreement with this work and the mean values of the predicted f_{esc} also reproduce the trend found there.

5. Summary

In this study, we see that the error in predicting the f_{esc} is comparable to the value in Jensen *et al.* (2016) where the spectra from a particular redshift was used. Thus, the prediction of the f_{esc} has weak dependence on the error we make in *unshifting* (predicting the redshift) the spectra. The spread on the prediction of f_{esc} is large. Jensen *et al.* (2016) showed that the spread will decrease by increasing the number of galaxies. For the real observations, we will be severely limited by quantity of data. However, even with these small sample sizes we will be able to extract useful information with our method. In our future paper (Giri *et al.*, in preparation), we will show that the method can be used to identify high LyC leakers.

References

Castellano, M., Amorín, R., Merlin, E., Fontana, A., McLure, R. J., Mármol-Queraltó, E., Mortlock, A., Parsa, S., Dunlop, J. S., Elbaz, D., *et al.* The astrodeep frontier fields catalogues-ii.

https://doi.org/10.1017/S1743921317011322 Published online by Cambridge University Press

photometric redshifts and rest frame properties in abell-2744 and macs-j0416. *Astronomy & Astrophysics*, 590: A31, 2016.

Ferland, G. J., Porter, R. L., Van Hoof, P. A. M., Williams, R. J. R., Abel, N. P., Lykins, M. L., Gargi, Shaw, Henney, W. J., & Stancil, P. C., The 2013 release of cloudy. *Revista mexicana de astronomía y astrofísica*, 49 (1): 137–163, 2013.

Gnedin, Nickolay Y. & Kaurov, Alexander A. Cosmic reionization on computers. ii. reionization history and its back-reaction on early galaxies. *The Astrophysical Journal*, 793 (1): 30, 2014.

Gunn, James E. & Peterson, Bruce A. On the density of neutral hydrogen in intergalactic space. *The Astrophysical Journal*, 142: 1633–1641, 1965.

Hartley, B. & Ricotti, M. Modelling reionization in a bursty universe. *MNRAS*, 462: 1164–1179, October 2016.

Ivezić, Željko., Connolly, Andrew J., VanderPlas, Jacob T., & Gray, Alexander. *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton University Press, 2014.

Izotov, Y. I., Orlitová, I., Schaerer, D., Thuan, T. X., Verhamme, A., Guseva, N. G., & Worseck, G. Eight per cent leakage of Lyman continuum photons from a compact, star-forming dwarf galaxy. *Nature*, 529: 178–180, January 2016.

Jensen, H., Zackrisson, E., Pelckmans, K., Binggeli, C., Ausmees, K., & Lundholm, U. A Machine-learning Approach to Measuring the Escape of Ionizing Radiation from Galaxies in the Reionization Epoch. *ApJ*, 827: 5, August 2016.

Jones, T. A., Ellis, R. S., Schenker, M. A., & Stark, D. P. Keck Spectroscopy of Gravitationally Lensed z ˜= 4 Galaxies: Improved Constraints on the Escape Fraction of Ionizing Photons. *ApJ*, 779: 52, December 2013.

Kohavi, Ron *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.

Leethochawalit, N., Jones, T. A., Ellis, R. S., Stark, D. P., & Zitrin, A. Absorption-line Spectroscopy of Gravitationally Lensed Galaxies: Further Constraints on the Escape Fraction of Ionizing Photons at High Redshift. *ApJ*, 831: 152, November 2016.

Lotz, J. M., Koekemoer, A., Coe, D., Grogin, N., Capak, P., Mack, J., Anderson, J., Avila, R., Barker, E. A., Borncamp, D., Brammer, G., Durbin, M., Gunning, H., Hilbert, B., Jenkner, H., Khandrika, H., Levay, Z., Lucas, R. A., MacKenty, J., Ogaz, S., Porterfield, B., Reid, N., Robberto, M., Royle, P., Smith, L. J., Storrie-Lombardi, L. J., Sunnquist, B., Surace, J., Taylor, D. C., Williams, R., Bullock, J., Dickinson, M., Finkelstein, S., Natarajan, P., Richard, J., Robertson, B., Tumlinson, J., Zitrin, A., Flanagan, K., Sembach, K., Soifer, B. T., & Mountain, M. The Frontier Fields: Survey Design and Initial Results. *ApJ*, 837: 97, March 2017.

Mitra, S., Choudhury, T. R., & Ferrara, A. Cosmic reionization after Planck. *MNRAS*, 454: L76–L80, November 2015.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Vanzella, E., de Barros, S., Vasei, K., Alavi, A., Giavalisco, M., Siana, B., Grazian, A., Hasinger, G., Suh, H., Cappelluti, N., Vito, F., Amorin, R., Balestra, I., Brusa, M., Calura, F., Castellano, M., Comastri, A., Fontana, A., Gilli, R., Mignoli, M., Pentericci, L., Vignali, C., & Zamorani, G. Hubble Imaging of the Ionizing Radiation from a Star-forming Galaxy at Z=3.2 with fesc>50%. *ApJ*, 825: 41, July 2016.

Zackrisson, E., Rydberg, C.-E., Schaerer, D., Östlin, G., & Tuli, M. The Spectral Evolution of the First Galaxies. I. James Webb Space Telescope Detection Limits and Color Criteria for Population III Galaxies. *ApJ*, 740: 13, October 2011.

Zackrisson, E., Inoue, A. K., & Jensen, H. The Spectral Evolution of the First Galaxies. II. Spectral Signatures of Lyman Continuum Leakage from Galaxies in the Reionization Epoch. *ApJ*, 777: 39, November 2013.

Zackrisson, E., Binggeli, C., Finlator, K., Gnedin, N. Y., Paardekooper, J.-P., Shimizu, I., Inoue, A. K., Jensen, H., Micheva, G., Khochfar, S., & Dalla Vecchia, C. The Spectral Evolution of the First Galaxies. III. Simulated James Webb Space Telescope Spectra of Reionization-epoch Galaxies with Lyman-continuum Leakage. *ApJ*, 836: 78, February 2017.