Bayesian Redshift Classification of Emission-line Galaxies with Photometric Equivalent Widths

Andrew S. Leung1,2, Viviana Acquaviva3, Eric Gawiser1, Robin Ciardullo1,4,5, Eiichiro Komatsu6,7, A. I. Malz8, Gregory R. Zeimann2,4,5, Joanna S. Bridge4,5, Niv Drory9, John J. Feldmeier10, Steven L. Finkelstein2, Karl Gebhardt2, Caryl Gronwall4,5, Alex Hagen4,5, Gary J. Hill2,9, and Donald P. Schneider

1 Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
2 Department of Astronomy, The University of Texas at Austin, Austin, TX 78712, USA; leung@astro.as.utexas.edu, gawiser@physics.rutgers.edu
3 Department of Physics, New York City College of Technology, The City University of New York, Brooklyn, NY 11201, USA; vacquaviva@citytech.cuny.edu
4 Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, PA 16802, USA
5 Institute for Gravitational and the Cosmos, The Pennsylvania State University, University Park, PA 16802, USA
6 Max-Planck-Institut für Astrophysik, D-85740 Garching bei München, Germany
7 Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU, WPI), Todai Institutes for Advanced Study, The University of Tokyo, Kashiwa 277-8583, Japan
8 Center for Cosmology and Particle Physics, Department of Physics, New York University, NY 10003, USA
9 McDonald Observatory, The University of Texas at Austin, Austin, TX 78712, USA
10 Department of Physics and Astronomy, Youngstown State University, Youngstown, OH 44555, USA

Received 2015 October 12; revised 2017 March 21; accepted 2017 May 4; published 2017 July 13

Abstract

We present a Bayesian approach to the redshift classification of emission-line galaxies when only a single emission line is detected spectroscopically. We consider the case of surveys for high-redshift Hα-emitting galaxies (LAEs), which have traditionally been classified via an inferred rest-frame equivalent width (EW; W_Hα) greater than 20 Å. Our Bayesian method relies on known prior probabilities in measured emission-line luminosity functions and EW distributions for the galaxy populations, and returns the probability that an object in question is an LAE given the characteristics observed. This approach will be directly relevant for the Hobby–Eberly Telescope Dark Energy Experiment (HETDEX), which seeks to classify ∼10^6 emission-line galaxies into LAEs and low-redshift [O II] emitters. For a simulated HETDEX catalog with realistic measurement noise, our Bayesian method recovers 86% of LAEs missed by the traditional W_Hα > 20 Å cutoff over 2 < z < 3, outperforming the EW cut in both contamination and incompleteness. This is due to the method’s ability to trade off between the two types of binary classification error by adjusting the stringency of the probability requirement for classifying an observed object as an LAE. In our simulations of HETDEX, this method reduces the uncertainty in cosmological distance measurements by 14% with respect to the EW cut, equivalent to recovering 29% more cosmological information. Rather than using binary object labels, this method enables the use of classification probabilities in large-scale structure analyses. It can be applied to narrowband emission-line surveys as well as upcoming large spectroscopic surveys including Euclid and WFIRST.

Key words: cosmological parameters — cosmology: observations — galaxies: distances and redshifts — galaxies: high-redshift — methods: statistical — quasars: emission lines

1. Introduction

The Hobby–Eberly Telescope Dark Energy Experiment11 (HETDEX) will obtain redshifts for approximately a million Lyα-emitting galaxies (LAEs) in its upcoming wide-field survey to determine the local Hubble expansion rate, angular diameter distance (d_A), and the growth of structure in the 1.9 < z < 3.5 universe (Hill et al. 2008). To achieve its science goals, HETDEX requires an accurate classifier to identify LAEs and discard contaminants (primarily z < 0.5 [O II] emitters) from the statistical LAE sample.

Lyα is a spectral feature produced by the transition of neutral hydrogen atoms from the first excited state to the ground state (n = 2 → 1), with a rest-frame wavelength of 1216 Å. The Visible Integral-field Replicable Unit Spectrograph (VIRUS), currently being deployed at the Hobby–Eberly Telescope (HET), has a spectral range of 3500–5500 Å (Hill & HETDEX Collaboration 2016), making spectroscopic detections of Lyα possible between redshifts 1.88 < z < 3.52. This capability enables the primary science goal of the HETDEX survey, the measurement of the expansion history of the universe (Hill et al. 2008; Adams et al. 2011), by constraining cosmological parameters from the observed signature of baryon acoustic oscillations (BAO) in the power spectrum of LAE redshifts and positions (Blake & Glazebrook 2003; Hu & Haiman 2003; Seo & Eisenstein 2003; Koehler et al. 2007; Seo & Eisenstein 2007; Shoji et al. 2009). Lyα is usually the strongest line, therefore near the survey limit, where most of the targets lie, the VIRUS spectra will have only one detected emission line.

Another prominent emission-line feature in galaxy spectra is the [O II] doublet at ~ 3726 Å and ~3729 Å, a pair of atomic transitions for the decay of singly ionized oxygen. Galaxies with strong [O II] emission from z < 0.476 will also be identified by HETDEX as a single-line detection within its spectral range, as the [O II] λ3727 doublet is separated by 2.5 Å (rest frame; Osterbrock 1974) and cannot be resolved by the VIRUS instrument (spectral resolution: 5.7 Å). Distinguishing LAEs targeted by HETDEX from low-redshift [O II] emitters

11 http://www.hetdex.org
We explore a Bayesian approach, using the posterior odds as a classifier, to identify emission-line galaxies as a means to improve the quality of the cosmological sample of LAEs. Section 2 describes our methodology, including details of the simulation of a HETDEX catalog consisting of LAEs and [O II] emitters (Section 2.1) and an overview of the statistical framework for a Bayesian method that can be used to identify LAEs in a line flux-limited sample of emission-line galaxies (Section 2.2). Section 3 presents the results of Bayesian classification of LAEs in a simulated HETDEX catalog and quantifies the improvement over the EW method. Section 4 offers a discussion of our findings and their applications.

Throughout the present work, we assume a \( \Lambda \)CDM cosmology with \( H_0 = 70 \text{ km s}^{-1} \text{Mpc}^{-1} \), \( \Omega_m = 0.3 \), and \( \Omega_\Lambda = 0.7 \) (Komatsu et al. 2011; Planck Collaboration et al. 2016). All magnitudes are reported in the AB system (Oke & Gunn 1983).

2. Methodology

2.1. Simulated Catalog of Emission-line Galaxies

We simulate populations of LAEs and [O II] emitters on which to test the methods for galaxy classification. For the simulations, we specify a 300 deg\(^2\) survey area (the size of the HETDEX spring field, roughly two-thirds of the total survey area) and a 1/4.5 filling factor to mimic the design of the upcoming HETDEX survey (Hill & HETDEX Collaboration 2016).

2.1.1. Spectroscopic Survey Simulation

Gronwall et al. (2007), Ciardullo et al. (2012), and C. Gronwall et al. (2016, in preparation; hereafter Gr16) measured the luminosity functions for LAE populations at \( z = 2.1 \) and \( z = 3.1 \). Using the Gr16 luminosity functions, we simulate Ly\( \alpha \) line luminosities via Monte Carlo simulations. We use a Schechter (1976) function of the form

\[
\Phi(L)dL = \phi^*(L/L^*)^{\alpha} \exp(-L/L^*)d(L/L^*),
\]

and assume that the parameters (shown in Table 1) evolve linearly with redshift. We obtain the distribution parameters at redshifts \( 2.1 < z < 3.1 \) by linear interpolation of the Gr16 parameter values for \( z = 2.1 \) and \( z = 3.1 \); for simulated LAEs at \( z < 2.1 \) and \( z > 3.1 \), we linearly extrapolate the parameters.

The top-left panel in Figure 2 shows that our extrapolation of the Gr16 luminosity function to \( z = 3.5 \) is consistent with the weakly evolving Ly\( \alpha \) luminosity functions measured at higher redshift (Shimasaku et al. 2006; Ouchi et al. 2008; Henry et al. 2012).

The measured EW distributions of LAEs have been modeled with various distributions. For our simulations, we assume an exponential form with a scale length, \( w_o \):

\[
\Psi(W)dW = \frac{e^{-W/w_o}}{w_o} dW.
\]

As with the Schechter function parameters, we assume the exponential scale length evolve linearly with redshift. Since previous studies have found weak or no correlation between emission-line luminosity and EW (Cowie et al. 1996; Hogg et al. 1998; Gronwall et al. 2007; Ciardullo et al. 2012, 2013), we model the luminosity function and the EW distribution as statistically independent.
The simulation of [O ii] luminosities follows the same procedure as described above, with Schechter function parameters reported by Ciardullo et al. (2013, hereafter Ci13). The [O ii] luminosity function has also been measured in studies by Gallego et al. (2002), Teplitz et al. (2003), Hippelein et al. (2003), Ly et al. (2007), Takahashi et al. (2007), and Comparat et al. (2015). For simplicity, we assume the exponential form in Equation (2) for the [O ii] EW distributions, with Ci13 parameters linearly evolving with redshift.

To model the evolution of the galaxy property distributions, we choose small values of $\Delta z$ for our simulations ($\Delta z = 0.01$ for LAEs; and $\Delta z = 0.005$ for [O ii] emitters, whose corresponding volume elements are smaller). The total number of each type of object to be simulated in a given redshift bin is given by the product of the comoving volume of the survey area in $\Delta z$ and the integral of the Schechter function above a conservative minimum flux, below which there can be no line detections on VIRUS. For each simulated object, the realized redshift and object type (LAE or [O ii] emitter) determine the observed-frame wavelength ($\lambda_{\text{obs}}$) of the primary simulated emission line. Since Ly$\alpha$ luminosity functions were measured for LAEs selected with the $W_{\text{Ly}\alpha} > 20$ Å requirement (e.g., Ciardullo et al. 2012; C. Gronwall et al. 2016, in preparation) and our Monte Carlo simulations realize EWs from probability distributions that go down to 0 Å, we must correct the number density of LAEs to account for the fraction that would have $W_{\text{Ly}\alpha} \leq 20$ Å.

Other emission lines are added to the simulated spectra of [O ii] emitters, based on relative line strengths as a function of metallicity (Anders & Fritze-v. Alvensleben 2003), which we assume to be one-fifth of solar for low-redshift objects in our simulations. The nearest [O ii] emitters ($z < 0.1$) have four other strong emission lines—[Ne iii] $\lambda$3869, H$\beta$ $\lambda$4861, [O iii] $\lambda$4959, and [O iii] $\lambda$5007—that may be detected within the spectral range of HETDEX (see Figure 1), depending on the redshift.

Gaussian noise with a mean equal to zero and a standard deviation equal to one-fifth of the wavelength-dependent 5σ line flux sensitivity limit of HETDEX is added to simulated line fluxes. Subsequent to the addition of noise, simulated objects with “recorded” line fluxes that fall below the 5σ detection limit are eliminated from the “observable” sample, resulting in a 5σ line flux-limited sample of emission-line galaxies for our simulated catalog.

2.1.2. Imaging Survey Simulation

We explore a scenario specific to our application (HETDEX), where spectroscopic line fluxes are coupled with continuum flux densities obtained through aperture photometry on broadband imaging, resulting in measurements of photometric EWs. Noiseless simulated emission-line fluxes ($f_n$) and observed-frame EWs ($W_{\text{obs}}$) are converted into continuum flux densities ($f_{\text{c,cont}}$) at the observed emission-line wavelength, as follows:

$$W_{\text{obs}} = \frac{f_n}{f_{\text{c,cont}}},$$

$$f_{\text{c,cont}} (\lambda_{\text{EL}}) = f_{\text{c,cont}} \lambda_{\text{EL}}^2 \frac{c}{\mu}.$$  

For each galaxy in the 5σ line flux-limited sample, we simulate an imaging survey counterpart by extrapolating its continuum flux density from the observed emission-line wavelength ($\lambda_{\text{EL}}$) to a specified broadband imaging survey filter. The power-law slope of the continuum is simulated from the distributions of $i'-z'$ colors of LAEs and [O ii] emitters observed in the HETDEX Pilot Survey13 (HPS; Adams et al. 2011; Blanc et al. 2011; Bridge et al. 2015).

\footnote{A Gaussian fit of the $i'-z'$ colors of LAEs in the HPS sample has a mean $\mu = 0.11$ and standard deviation $\sigma = 0.55$; for [O ii] emitters in HPS, a Gaussian fit of $i'-z'$ has $\mu = 0.53$ and $\sigma = 0.28$.}

Notes.

\footnote{Log-normal distribution parameter $\sigma_y$ is fixed at 1.1 for all redshifts $z_{\text{LAE}}$; see Equation (29).}

\footnote{The consequences of using log-normally distributed [O ii] EWs are discussed in Section 4.1.}

\footnote{C. Gronwall et al. (2016, in preparation).}

\footnote{Ciardullo et al. (2012).}

\footnote{Gronwall et al. (2007).}

\footnote{Ciardullo et al. (2013).}
The Astrophysical Journal, 843:130 (15pp), 2017 July 10

Leung et al.

Figure 2. Luminosity functions and equivalent width distributions for LAEs and \([\text{O} \, \text{II}]\) emitters. Upper left: dashed black curve is obtained by linear extrapolation of Schechter function parameters reported by C. Gronwall et al. (2016, in preparation) out to \(z = 3.5\); it is consistent with \(L_{\text{Ly} \alpha}\) luminosity functions measured at higher redshifts (Shimasaku et al. 2006; Ouchi et al. 2008; Henry et al. 2012). Upper right: \(L_{\text{Ly} \alpha}\) EW distributions assumed in simulations in the present work; C. Gronwall et al. (2016, in preparation) parameters evolving linearly with redshift constitute our “baseline” scenario (see Table 4). Lower right: dashed lines show log-normal fits at \(\text{EW} > 5 \text{ Å}\) to the exponential equivalent width distributions measured by Ciardullo et al. (2013; see Section 4.1); the exponential form is used in simulations of \([\text{O} \, \text{II}]\) equivalent widths in the present work. Lower left: \([\text{O} \, \text{II}]\) luminosity functions used in our simulations (Ciardullo et al. 2013) are consistent with published results at similar redshifts (Gallego et al. 2002; Comparat et al. 2015).

procedure prescribed by Madau (1995) is applied to simulated spectra as a function of observed wavelength to correct for absorption by neutral hydrogen in the intergalactic medium (the “Ly\(\alpha\) forest”; Lynds 1971). The sum of the resulting flux from the continuum and flux contributed by emission lines, multiplied by the transmission fraction of the specified imaging filter (including the quantum efficiency of the CCD) results in the continuum flux density observed for each galaxy in simulated aperture photometry. One fifth of the \(5\sigma\) depth of the simulated imaging survey (SDSS \(g'\) and \(r'\) filters are assumed in this work; Doi et al. 2010) is used as the standard deviation for the Gaussian profile of measurement noise in photometry.

For each model \(5\sigma\) emission-line detection, we compute a new quantity—photometric EW—as the relative strength of the simulated line flux to the continuum flux density measured in aperture photometry. These simulated photometric EWs are observed-frame quantities with measurement noise propagated from both simulated line flux and simulated continuum flux density measurements. As a result of their larger EWs and greater luminosity distances from Earth, among \(5\sigma\) line detections LAEs are generally fainter than \([\text{O} \, \text{II}]\) emitters in their continua, and their measurements therefore suffer from larger fractional uncertainties than continuum measurements of \([\text{O} \, \text{II}]\) emitters.

Figure 3 presents the distribution of inferred \(\text{Ly}\alpha\) EWs (\(W_{\text{Ly} \alpha}\)) for a \(5\sigma\) line flux-limited sample of simulated LAEs and corresponding \([\text{O} \, \text{II}]\) emitters, plotted against their continuum flux densities \((f_{\text{cont}})\) obtained in simulated aperture photometry on \(g'\) band imaging with \(5\sigma\) depth of 25.1 mag. The horizontal dashed line denotes an EW of 20 Å in the rest frame of \(\text{Ly}\alpha\) emission.

2.2. Bayesian Classification of Emission-line Galaxies

HETDEX will use the two-point correlation function, or power spectrum, of high-redshift LAEs to measure dark energy evolution over \(1.9 < z < 3.5\) (Hill et al. 2008; Adams et al. 2011). When \([\text{O} \, \text{II}]\) emitters are misidentified as LAEs, the correlations of low-redshift galaxies with smaller comoving separations are erroneously mapped to the correlation function of LAEs (Komatsu in Appendix). The observed power spectrum is given by the weighted average of the LAE power spectrum, \(P_{\text{LAE}}\), and the contamination power spectrum,

\[
P_{\text{obs}}(k_\perp, k) = (1 - f_{\text{[O II]}})^2 P_{\text{LAE}}(\sqrt{k^2 + k_\perp^2}) + f_{\text{[O II]}}^2 (\alpha^2/\beta) P_{\text{[O II]}}(\sqrt{\alpha^2 k_\perp^2 + \beta^2 k^2}), \tag{5}
\]

where \(\alpha\) and \(\beta\) are parameters that describe the way in which the observed power spectrum is contaminated by signal from

---

\(^{14}\) See the Appendix for the derivation of Equation (5).
low-redshift [O II] emitters in the survey foreground (see Equation (36) for the definition of the contamination power spectrum and the footnotes of Table 5 for the definitions of $\alpha$ and $\beta$), and $f_{[O\ II]}$ is the fraction of [O II] emitters in the total LAE sample, i.e.,

$$f_{[O\ II]} = \frac{\text{number of contaminating [O II] emitters}}{\text{number of galaxies classified as LAEs}}. \quad (6)$$

Improvement in the quality of the LAE sample can be achieved by increasing the completeness of the statistical sample and/or reducing contamination by [O II] emitters relative to the sample obtained by the minimum EW requirement. A Bayesian method takes other observed information into consideration in addition to the EW for each 5σ line detection for the purpose of making a classification. This additional information comprises the wavelength at which the detected emission line is observed ($\lambda_{\text{det}}$), the flux measured for the targeted emission line, and the detection (or non-detection) of other emission lines expected to fall within the spectral range of HETDEX if the emission line is [O II].

2.2.1. Bayes’ Theorem and Classification Threshold

Bayes’ theorem gives the posterior probability of a hypothesis or model $H$ given the observed data $B$,

$$P(H|B) = \frac{P(B|H)P(H)}{P(B)}. \quad (7)$$

For a discrete set of $N$ models, the probability of the data $B$ is simply

$$P(B) = \frac{1}{N} \sum_{i=1}^{N} P(B|H_i). \quad (8)$$

For two competing models $H_1$ and $H_2$, the posterior odds ratio is

$$\frac{P(H_1|B)}{P(H_2|B)} = \frac{P(H_1)}{P(H_2)} \frac{P(B|H_1)}{P(B|H_2)}, \quad (9)$$

where the first term on the right-hand side is the prior odds ratio and the second term, the ratio of the marginal likelihoods under the two models, is called the Bayes factor (Gelman et al. 1995; Ivezić et al. 2014). The normalization given by Equation (8) cancels out.

For our application in the present work, in which we seek to classify galaxies detected via a single emission line into two samples, namely, high-redshift LAEs (the target sample) and foreground [O II] emitters (the dominant contaminant), the relevant posterior odds ratio is

$$\frac{P(\text{LAE}|\text{data})}{P(\text{[O II]}|\text{data})} = \frac{P(\text{LAE})P(\text{data}|\text{LAE})}{P(\text{[O II]}|\text{data})P(\text{data}|\text{[O II]})}. \quad (10)$$

The natural threshold of the posterior odds ratio to classify an emission-line detection as LAE is $>1$, but this threshold parameter may be tuned to optimize cosmological results, that is, to minimize variance in $d_A$, our estimator of the two-point correlation function.

2.2.2. Prior Probabilities and Prior Odds Ratio

The prior probability that an object is an LAE or an [O II] emitter is calculated as a function of wavelength, using the luminosity functions and cosmological volume elements at the corresponding redshifts. For a given model, an observed emission-line wavelength $\lambda_{\text{el}}$ corresponds to exactly one redshift $z$,

$$\lambda_{\text{el}} = (1 + z) \lambda, \quad (11)$$

where $\lambda$ is the line wavelength in the emission rest frame. The assumed cosmology (given in Section 1) determines the comoving volume $\delta V_c$ of each redshift “slice” $\delta z$, corresponding to a wavelength interval $\delta \lambda$ in the observed frame.\textsuperscript{15}

Under the simplifying assumption that there are only two types of emission-line objects, the prior probability that a detected line is Lyα and the prior probability that it is [O II] sum up to unity for a given interval $\delta \lambda$,

$$P(\text{LAE}) + P(\text{[O II]}) = 1. \quad (12)$$

Therefore, the prior probability that a detected line is Lyα is the product of the number density $n$ at redshift $z$ and the differential comoving volume at a corresponding redshift interval $\delta z$, given as a fraction of the total number of objects expected to populate the differential volumes associated with LAEs and [O II] emitters for the observed emission-line wavelength,

$$P(\text{LAE}) = \frac{n_{\text{LAE}} \delta V_{\text{LAE}}}{n_{\text{LAE}} \delta V_{\text{LAE}} + n_{\text{[O II]}} \delta V_{\text{[O II]}}}. \quad (13)$$

\textsuperscript{15} Here, $\delta \lambda$ denotes a fixed interval of $\lambda_{\text{el}}$ and not the uncertainty of each detected line.
Similarly, the prior probability that a detected line is \([\text{O} \text{ II}]\) is

\[
P([\text{O} \text{ II}]) = \frac{n_{\text{LAE}} \delta V_{\text{LAE}}}{n_{\text{O II}} \delta V_{\text{O II}} + n_{\text{LAE}} \delta V_{\text{LAE}}}. \tag{14}
\]

The volume of the interval \(z \pm \delta z\) is given by the comoving volume integrated over the redshift interval corresponding to a wavelength interval around the observed emission-line wavelength, \(\lambda_{\text{el}} \pm \delta \lambda\),

\[
\delta V_{\text{LAE}}(\lambda_{\text{el}}) = \delta V(z_{\text{LAE}}) = \int_{z_{\text{LAE}}+\delta z}^{z_{\text{LAE}}+\delta z} dV_c, \tag{15a}
\]

\[
\delta V_{\text{O II}}(\lambda_{\text{el}}) = \delta V(z_{\text{O II}}) = \int_{z_{\text{O II}}+\delta z}^{z_{\text{O II}}+\delta z} dV_c. \tag{15b}
\]

The integral of the Schechter function (Equation (1)) down to the line luminosity corresponding to the flux detection limit yields the number densities of LAEs and \([\text{O II}]\) emitters:

\[
n_{\text{LAE}}(\lambda_{\text{el}}) = \int_{\lambda_{\text{LAE}}}^{\infty} \phi_{\text{LAE}} (L/L_{*})^{\alpha_{\text{LAE}}} e^{-L/L_{*}} dL, \tag{16a}
\]

\[
n_{\text{O II}}(\lambda_{\text{el}}) = \int_{\lambda_{\text{O II}}}^{\infty} \phi_{\text{O II}} (L/L_{*})^{\alpha_{\text{O II}}} e^{-L/L_{*}} dL, \tag{16b}
\]

where the vector of parameters \(\theta = (\alpha, L_{*}, \phi_{*})\) as given in Table 1, and the lower limits of integration are related to the wavelength-dependent flux detection limit \(f_{\text{min}}\) as follows:

\[
L_{\text{min}}(\lambda_{\text{el}}) = 4\pi [d_L(z_{\text{LAE}})]^2 f_{\text{min}}(\lambda_{\text{el}}), \tag{17a}
\]

\[
L_{\text{O II}}(\lambda_{\text{el}}) = 4\pi [d_L(z_{\text{O II}})]^2 f_{\text{min}}(\lambda_{\text{el}}), \tag{17b}
\]

where \(d_L\) is luminosity distance and \(z = z(\lambda_{\text{el}})\) as defined in Equation (11).

The prior odds ratio is therefore

\[
P([\text{LAES}) = \frac{n_{\text{O II}}(\lambda_{\text{el}}) \delta V_{\text{O II}}(\lambda_{\text{el}})}{n_{\text{LAE}}(\lambda_{\text{el}}) \delta V_{\text{LAE}}(\lambda_{\text{el}})}. \tag{18}
\]

In the case of HETDEX, the assumption of Equation (12) is reasonable. Since the VIRUS spectrographs only extend to 5500 Å, lines such as H\(\beta\) and the \([\text{O III}]\) doublet will not be observable past \(z \sim 0.13\) (see Section 3.3). Moreover, while AGNs may produce strong line emission at 1549 Å (C IV), 1909 Å (C III), and 2798 Å (Mg II), the VIRUS coverage is such that these objects will seldom be single-line detections (see Section 4.5). By far, the dominant source of confusion is between \([\text{O II}]\) and Ly\(\alpha\) (Adams et al. 2011).

16 Note that \(\delta z\), which corresponds to a fixed interval \(\delta \lambda\) in the observed frame (Equation (11)), has different values under the assumption of each model:

\[
z_{\text{LAE}} \pm \delta z = (\lambda_{\text{el}} \pm \delta \lambda)/1216 \text{ Å} - 1,
\]

\[
z_{\text{O II}} \pm \delta z = (\lambda_{\text{el}} \pm \delta \lambda)/3727 \text{ Å} - 1.
\]

17 The Schechter function integral down to a specified lower limit is the upper incomplete gamma function:

\[
\Gamma(x, s) = \int_{s}^{\infty} t^{s-1} e^{-t} dt.
\]

The luminosity functions and EW distributions reported by C. Gronwall et al. (2016, in preparation) for \(z = 2.1\) and \(z = 3.1\) populations of LAEs represent the best current information on the properties of LAEs in the HETDEX redshift range. These distributions take the form of probability density functions: \(\Psi(L)\) gives a number density per unit volume as a function of intrinsic line luminosity and \(\Psi(W)\) is a normalized probability density as a function of EW in the emission rest frame. Given these distributions, we can calculate the conditional probability of measuring the observed data under the assumption that the object is an LAE. The conditional probability of the observation given that the observed object is an \([\text{O II}]\) emitter is similarly obtained from the galaxy property distributions for \([\text{O II}]\) emitters sampled at \(0 < z < 0.56\) by Ciardullo et al. (2013). Since any correlation between emission-line luminosity and EW is, at best, weak (Cowie et al. 1996; Hogg et al. 1998; Gronwall et al. 2007; Ciardullo et al. 2012, 2013), we treat these two quantities as statistically independent.

The likelihood function of the observables \(B\) under the assumption of model \(H_j\) is

\[
P(B|H_j) = P(f_{\text{el}}(\lambda_{\text{el}}), \text{EW}_{\text{obs}}(\theta_j))
\]

\[
= \frac{1}{n_j} \int_{a_{\text{el}}}^{a_{\text{el}}} \phi_j(L) dL \int_{b_{\text{el}}}^{b_{\text{el}}} \Psi_j(W) dW, \tag{19}
\]

where \(j \in \text{[LAE, [O II]]}\); the vector of parameters \(\theta_j = (a_j, L_{*j}, \phi_{*j}, w_{j})\); \(n_j\) is the normalization of the Schechter function integral as given in Equations (16); the luminosity function \(\Phi(L)\) and the EW distribution \(\Psi(W)\) are given in Equations (1) and (2), respectively; and the limits of integration are

\[
a_{\text{el}} = L_{\text{el}}(\lambda_{\text{el}}), f_{\text{el}} \pm \delta L_{\text{el}}, \tag{20}
\]

\[
b_{\text{el}} = W_{\text{el}}(\lambda_{\text{el}}, \text{EW}_{\text{obs}}) \pm \delta W_{\text{el}}, \tag{21}
\]

where \(j \in \text{[LAE, [O II]]}\); \(\delta L_{\text{el}}\) and \(\delta W_{\text{el}}\) denote differential changes to the corresponding model quantities, with the fractional amount of change held fixed across the hypotheses. The exponential form of the EW distribution \(\Psi(W) dW\) given in Equation (2) already includes the proper normalization.\(^{19}\)

The luminosity of the emission line is calculated as in Equations (17), but with \(f_{\text{el}}\) replacing \(f_{\text{min}}\) i.e.,

\[
L_{\text{LAE}}(\lambda_{\text{el}}, f_{\text{el}}) = 4\pi [d_L(z_{\text{LAE}}(\lambda_{\text{el}}))]^2 f_{\text{el}}, \tag{22a}
\]

\[
L_{\text{O II}}(\lambda_{\text{el}}, f_{\text{el}}) = 4\pi [d_L(z_{\text{O II}}(\lambda_{\text{el}}))]^2 f_{\text{el}}. \tag{22b}
\]

The EW of the emission line in the rest frame is related to its observed EW by a factor of \(1 + z\):

\[
W_{\text{LAE}}(\lambda_{\text{el}}, \text{EW}_{\text{obs}}) = \frac{\text{EW}_{\text{obs}}}{1 + z_{\text{LAE}}} = \text{EW}_{\text{obs}} \frac{1216 \text{ Å}}{\lambda_{\text{el}}}, \tag{23a}
\]

\[
W_{\text{O II}}(\lambda_{\text{el}}, \text{EW}_{\text{obs}}) = \frac{\text{EW}_{\text{obs}}}{1 + z_{\text{O II}}} = \text{EW}_{\text{obs}} \frac{3727 \text{ Å}}{\lambda_{\text{el}}}. \tag{23b}
\]

\(^{19}\) The same is true for the log-normal form to be given in Equation (29) in Section 4.1.
The Bayes factor is therefore

$$\frac{P(\text{data}|\text{LAE})}{P(\text{data}|[\text{O} \ II])} = \frac{n_{[\text{O} \ II]}^{-} \int_{L_{[\text{O} \ II]}}^{L_{\text{LAE}}} \Phi_{[\text{O} \ II]}(L) dL \int_{W_{[\text{O} \ II]}}^{W_{\text{LAE}}} \Psi_{[\text{O} \ II]}(W) dW}{n_{\text{LAE}}^{-} \int_{L_{[\text{O} \ II]}}^{L_{\text{LAE}}} \Phi_{\text{LAE}}(L) dL \int_{W_{\text{LAE}}}^{W_{\text{LAE}}} \Psi_{\text{LAE}}(W) dW}.$$  \hspace{1cm} (24)

By combining the above with Equation (18), the posterior odds ratio, our Bayesian classifier introduced in Equation (10), evaluates to

$$\frac{P(\text{LAE}|\text{data})}{P([\text{O} \ II]|\text{data})} = \frac{\delta V_{\text{LAE}}}{{\delta V_{[\text{O} \ II]}}} \int_{L_{[\text{O} \ II]}}^{L_{\text{LAE}}} \Phi_{[\text{O} \ II]}(L) dL \int_{W_{[\text{O} \ II]}}^{W_{\text{LAE}}} \Psi_{[\text{O} \ II]}(W) dW \frac{\delta V_{[\text{O} \ II]}}{\delta V_{\text{LAE}}} \int_{L_{\text{LAE}}}^{L_{\text{LAE}}} \Phi_{\text{LAE}}(L) dL \int_{W_{[\text{O} \ II]}}^{W_{\text{LAE}}} \Psi_{[\text{O} \ II]}(W) dW. \hspace{1cm} (25)$$

The initial season of HETDEX data will measure the luminosity functions and EW distributions of both populations to high precision. In practice, HETDEX data will allow for iterative refinement of our Bayesian priors as the luminosity functions and EW distributions of LAEs and [O II] emitters, and their evolution as functions of redshift, become more precisely known. We will discuss sensitivity to model assumptions and prior specification in Section 4.1.

3. Results

3.1. Cosmological Distance Measurement Uncertainties

In order to obtain an indicator of the performance of each classification method, we parameterize the fractional uncertainty in angular diameter distance ($d_A$) measurements as a function of contamination and incompleteness of the statistical sample of LAEs, as follows:

$$\frac{\sigma_{d_A}}{d_A} = A x_0^0 + C x_1 + D x_0^F + E x_0^S + F,$$  \hspace{1cm} (26)

where $x_0 = f_{[\text{O} \ II]}$ is fractional contamination (as defined in Equation (6)) and $x_1 = 1 - N_{\text{true LAEs}} / N_{\text{available}}$ is sample incompleteness. We use the observable power spectrum (Equation (5)) in a Fisher (1935) matrix code\textsuperscript{20} (Shoji et al. 2009) that marginalizes over the contamination power spectrum (the second term in Equation (5)) to obtain $\sigma_{d_A}/d_A$ for a grid of contamination and incompleteness values. A linear bias factor of 2.0 is used for LAEs in both redshift bins. We then use the grid of results to derive a fitting formula for the parameters. The uncertainty parameter, $\sigma_{d_A}/d_A$, for Ly α emitters in the simulated data is given by the parameter $H$ in Table 2.

Results for Bayesian classification of emission-line galaxies in a simulated HETDEX catalog are presented for an optimized requirement of the posterior odds ratio (Equation (25)) for selection as LAEs. This requirement minimizes $\sigma_{d_A}/d_A$ given perfect information on the simulated object labels by optimizing the trade off between contamination and incompleteness.

Each redshift bin may be optimized independently to maximize the total amount of information obtained from the full $1.9 < z < 3.5$ LAE sample. To accomplish this goal, we need to determine a set of values for the eight parameters in Equation (26) for each redshift bin. In our analysis we divided the full spectral range of HETDEX into two redshift bins and tuned the required posterior odds ratio for LAE classification in each bin separately; Figure 4 and Tables 3 and 4 present results corresponding to a Bayesian method optimized separately for two redshift bins. The value of $\sigma_{d_A}/d_A$ in total is estimated by inverse variance summation.

3.2. Improvement over Traditional EW Method

Compared to the traditional $W_{\lambda, 20}$ > 20 Å narrowband limit to classify emission-line galaxies as LAEs (which discards all data below the dashed line in Figure 3), the Bayesian method presented in Section 2.2 recovers a more complete statistical sample of high-redshift LAEs without an overall increase in misidentified low-redshift [O II] emitters. Our Bayesian method is adaptive to prior probabilities that reflect the evolution of the galaxy populations and the effect of cosmological volume on the relative density of galaxies as a function of wavelength.

LAEs at $z < 2.065$ are not contaminated by foreground [O II] emitters, since [O II] will not be detected at $\lambda_{\text{LS}} < 3727$ Å. Moreover, at $z < 0.05$, a galaxy’s angular scale is greater than 1” per kiloparsec, hence all but the most compact [O II] sources will be resolved in the imaging survey. Consequently, out to about $z \sim 2.4$, our Bayesian analysis recovers all LAEs with negligible [O II] contamination (see Figure 4, top row).

The rate of contamination in the LAE sample identified by our Bayesian method is sub-percent up to $z \sim 3.0$. Over $1.88 < z < 2.54$, the Bayesian method recovers more than 99% of available LAEs, compared to a sample identified by the traditional $W_{\lambda, 20} > 20$ Å cutoff that is only ∼70% complete.

Table 3 provides a comparative summary of the two classification methods. With respect to the traditional $W_{\lambda, 20} > 20$ Å cut, the Bayesian method significantly increases the completeness of the sample of objects classified as LAEs by trading near-zero contamination in the case of the EW method for sub-percent contamination in the low-redshift bin ($1.9 < z < 2.5$).

Over the entire HETDEX spectral range (3500–5500 Å), our Bayesian method recovers ∼25% more LAEs than the traditional EW method. Over the redshift range $2 < z < 3$,

\textsuperscript{20} http://wwwmpa.mpa-garching.mpg.de/~komatsu/crl/list-of-routines.html
Figure 4. Four selected redshift bins shown in rows for a simulated HETDEX survey with $g'$ ($5\sigma = 25.1$) band imaging survey and realistic measurement noise. Approximately 5% of the galaxies in each redshift bin are plotted. Left: all $5\sigma$ spectroscopic detections of emission-line galaxies whose primary lines are observed in the given wavelength interval. Middle (right): simulated emission-line galaxies classified by Bayesian method into high-redshift LAE (foreground [O II]) samples at the corresponding redshifts. Correctly classified "true" LAEs ([O II] emitters) are shown in red (blue); misidentified [O II] emitters (erroneously discarded LAEs) are indicated in blue (red).
Table 3

Classification Results for a Simulated HETDEX Catalog Based on \( z > 16 \) Luminosity Functions and Equivalent Width Distributions, with Simulated Aperture Photometry on \( g' (5σ = 25.1) \) Band Imaging

| Classification Method | \( W_{\text{Ly}α} > 20 \text{ Å} \) | Bayesian Method |
|-----------------------|-------------------------------|------------------|
| Galaxies classified as LAEs | 667,400 | 847,500 | 796,200 |
| Missed observable LAEs | 218,700 | 20,100 | 50,800 |
| Sample incompleteness | 26.0% | 2.39% | 6.02% |
| Misidentified \([\text{O} \, \text{II}]\) emitters | 13,500 | 25,100 | 4,400 |
| Fractional contamination | 2.12% | 2.96% | 0.55% |
| Measurement uncertainty, \( \sigma_{f}/d_{f} \) | 1.32% | 1.19% | 1.16% |

Notes.

\( a \) The “default” Bayesian requirement for LAE classification is \( P(\text{LAE} | \text{data}) ≥ 1 \).

\( b \) The “optimized” Bayesian method requires \( P(\text{LAE} | \text{data}) > 1.38 \) for the classification of \( 1.9 < z < 2.5 \) LAEs and \( P(\text{LAE} | \text{data}) > 10.3 \) for \( 2.5 < z < 3.5 \) LAEs.

86% of “true” LAEs missed by \( W_{\text{Ly}α} > 20 \text{ Å} \) are correctly classified by the Bayesian method, representing the recovery of cosmological information that would be discarded by the EW method. In addition to improving the completeness of the LAE sample, the Bayesian method also reduces contamination by \([\text{O} \, \text{II}]\) sources by a factor of \( \sim 4 \) in our simulations.

3.3. Aperture Spectroscopy for Additional Emission Lines

The presence of other emission lines redward of the primary detected line (in the case of \([\text{O} \, \text{II}]\) emitters; see Figure 1), when they fall within the spectral range of HETDEX, provides additional observed information for the two likelihood functions in the Bayes factor. Accounting for this spectral information leads to better classification performance by the Bayesian method in the form of additional reductions in fractional contamination and further increases in the completeness of the LAE sample.

At \( z < 0.1 \), the vast majority of \([\text{O} \, \text{II}]\)-emitting galaxies will be detected via multiple emission lines; we typically observe stronger \([\text{Ne} \, \text{III}] \lambda 5007 \) emission than \([\text{O} \, \text{II}] \lambda 3727 \). For the bulk of the HETDEX \([\text{O} \, \text{II}] \) redshift range, \( 0.13 < z < 0.42 \), \([\text{Ne} \, \text{III}] \lambda 3869 \) is the only other line that falls into the spectrograph’s bandpass.

With our previous assumption of statistically independent quantities, the likelihood functions \( P(\text{data} | \text{LAE}) \) and \( P(\text{data} | [\text{O} \, \text{II}]) \) are each the product of the likelihood functions associated with the individual properties we wish to consider:

\[
P(\text{data}|\text{type}) = \prod_{\text{properties}} P(\text{data}|\text{type})_{\text{property}},
\]

In particular,

\[
P(\text{data}|\text{type})_{\text{lines}} = \prod_{\text{lines}} P(\text{data}|\text{type})_{\text{line}},
\]

where the lines we wish to consider are \([\text{Ne} \, \text{III}] \lambda 3869, [Hβ] \lambda 4861, [\text{O} \, \text{III}] \lambda 4959, \) and \([\text{O} \, \text{III}] \lambda 5007 \), subject to their falling within the HETDEX spectral range. An example of other “properties” one might include in this analysis is the observed color of objects, obtained by having imaging data in more than one band (see discussion in Section 4.4).

Assuming a Gaussian noise distribution, we can calculate the probability of the measured flux at each expected line location. When a line is out of range, it contributes no information for or against the hypothesis that the primary detected line is \([\text{O} \, \text{II}]\); the \( P(\text{data}|\text{type})_{\text{line}} \) in question is set equal to unity, thereby having no effect on the value in the left-hand side of Equation (28).

The improvement due to accounting for additional emission lines is evident when we consider the boundary at which spectroscopic information from all additional lines is lost, when \([\text{Ne} \, \text{III}] \lambda 3869 \) is redshifted out of the HETDEX spectral range (3500–5500 Å) for \( z > 0.42 \) \([\text{O} \, \text{II}] \) emitters (\( \lambda_{\text{Ly}α} > 5299 \) Å). The bottom row in Figure 4 shows 5σ emission-line detections at \( 5300 < \lambda_{\text{Ly}α} < 5500 \) Å and their classification into samples of LAEs and \([\text{O} \, \text{II}]\) emitters. Without spectroscopic information from the additional lines, the Bayesian cutoff between LAEs and \([\text{O} \, \text{II}]\) emitters is reduced to a nearly straight line on a log–log plot of \( W_{\text{Ly}α} \) versus continuum flux densities \( (f_{\text{c,cont}}) \) in this redshift bin, which is the reddest 200 Å in the spectral range of HETDEX (compare to the third row in Figure 4).

3.4. Optimizing Area versus Depth in Fixed Observing Time

Using our Bayesian method and HETDEX as a baseline scenario, we investigate the survey design trade off between total survey area and depth of coverage per unit survey area. Holding the amount of available observing time fixed at the HETDEX survey design (denoted by the gray dashed line in Figure 5), we apply 5σ depths in both simulated spectroscopic and imaging surveys that are modified by \( 1/\sqrt{t} \), where \( t \) is the factor by which observing time per unit survey area changes as a result of a corresponding change in total survey area. Simulated measurement noise varies accordingly, as described in Section 2.1.

The number of LAEs available to be recovered in the 5σ line flux-limited sample changes with survey design, as shown in the upper panel of Figure 5. For each survey design, we re-run the Fisher matrix code described in Section 3.1 with the number of available LAEs to determine a new set of parameters for Equation (26) (i.e., Table 2) and re-optimize the Bayesian method for each case.

Our analysis indicates that the current HETDEX survey design is effectively optimal in the trade off between area and depth when our Bayesian method is used as the redshift classifier: trading away from the nominal 300 deg² survey area moves the optimal \( \sigma_{f}/d_{f} \) for the two redshift bins in opposite directions (lower panel in Figure 5).

4. Discussion

4.1. Imperfect Knowledge of Distribution Functions of Galaxy Properties

The luminosity functions and EW distributions measured by Ciardullo et al. (2013) and C. Gronwall et al. (2016, in preparation) represent the best current information on the galaxy populations HETDEX will observe, but will be supplemented by data collected in the initial season of HETDEX observations. The ability of the Bayesian method to classify spectroscopic emission-line detections does not crucially depend on perfect knowledge of the characteristics of the galaxy populations. To demonstrate this behavior, we test our
method with simulated populations of LAEs with varying characteristics, including cases in which the luminosity function and EW distribution do not evolve with redshift. For these tests, the population of foreground [O II] emitters is simulated as described in Section 2.1; the Bayesian method assumes the priors given in Section 2.2 and uses the LAE classification cutoff optimized for the “baseline” scenario (Table 3). Table 4 compares two such test cases with the “baseline” scenario.

The \( z = 3.1 \) Ly\( \alpha \) luminosity function measured by Gronwall et al. (2007), if assumed to be constant with redshift and applied to the entire redshift range \( (1.9 < z < 3.5) \), implies more observable LAEs (whose emission-line fluxes exceed the detection limit) at both low and high redshift (see upper left panel of Figure 2). This result leads to a lower rate of contamination in the LAE sample classified by a Bayesian method that uses the same priors as the baseline and which is optimized with respect to the baseline scenario. Although the Bayesian method recovers a larger LAE sample in this scenario, the sample is more incomplete in fractional terms due to the large number of “true” observable LAEs available to be recovered.

The \( z = 2.1 \) Ly\( \alpha \) luminosity function measured by Ciardullo et al. (2012), if assumed to be constant with redshift and applied to the entire redshift range, represents a scenario that is unfavorable for LAE-based cosmological study at high redshift. Nevertheless, our Bayesian method misses fewer observable LAEs and misidentifies fewer [O II] emitters in this unfavorable scenario than does the EW method on either counts in the sample of objects classified as LAEs under 1%, down to survey depth AB \( \sim 22 \) mag for the low-redshift bin \( (1.9 < z < 2.5) \) and AB \( \sim 24 \) mag for the high-redshift bin \( (2.5 < z < 3.5) \).

Figure 6 illustrates that for the 3500–4300 Å portion of the HETDEX spectral range (top panel), essentially all observed [O II] emitters have continuum flux densities greater than the 5\( \sigma \) limit of HETDEX broadband imaging, resulting in small EWs and enabling the success of the traditional method for LAE selection in this regime of observed emission-line wavelengths. The inclusion of emission-line flux in addition to EW in Bayesian classification leads to an LAE sample with a rate of contamination less than a quarter percent.

In the longer wavelength bin (4300–5500 Å, bottom panel of Figure 6), line fluxes measured for [O II] emitters are weaker than those in the shorter-wavelength bin due to increasing luminosity distance, rendering [O II] EWs measured in this bin similar to those of high-redshift LAEs. In this wavelength range, accurate classification is relatively difficult for both methods. Equivalent width alone is not an adequate classifier; inclusion of the emission-line flux of the primary detection and information from additional lines enables Bayesian classification to be effective in a regime that poses serious challenges to the traditional W\(_{\text{Ly}\alpha} > 20 \) Å method.

### 4.2. Sensitivity to Imaging Survey Depth

Since the vast majority of [O II] emitters are brighter in their continua than the imaging depth in the survey design of HETDEX, the Bayesian method is able to keep contamination in the sample of objects classified as LAEs under 1%, down to survey depth AB \( \sim 22 \) mag for the low-redshift bin \( (1.9 < z < 2.5) \) and AB \( \sim 24 \) mag for the high-redshift bin \( (2.5 < z < 3.5) \).

### 4.3. Single Broadband Filter with Best Performance

The performance of our Bayesian method is similar with \( g' \) and \( r' \) band simulated imaging in equal observing time,

---

**Table 4**

Bayesian Method Classification Results for Three Simulated HETDEX Catalogs with \( g' \) Band Imaging (5\( \sigma = 25.1 \)); “Baseline” Distributions are used as Bayesian Priors for Each Simulation Scenario

| Simulation Scenario | Baseline | Pessimistic | Optimistic |
|---------------------|----------|-------------|------------|
|                      | \( \text{Gr16} \) | \( \text{Cl12}, z = 2.1 \) | \( \text{Gr07}, z = 3.1 \) |
| Distribution of LAEs| Evolving | No Evolution | No Evolution |
| [O II] emitters      | \( \text{Cl13} \) | \(+1\{\phi^*, L^*, w_0\}_{\text{Cl13}}\) | \(-1\{\phi^*, L^*, w_0\}_{\text{Cl13}}\) |
| “True” observable LAEs | 842,600 | 375,000 | 1,155,000 |
| Galaxies classified as LAEs | 796,200 | 345,000 | 1,075,800 |
| Missed observable LAEs | 50,800 | 32,700 | 83,500 |
| Sample incompleteness | 6.02% | 8.73% | 7.23% |
| Misidentified [O II] emitters | 4420 | 7700 | 4350 |
| Fractional contamination | 0.55% | 2.20% | 0.40% |

---

Figure 2: the fitted parameters are presented in the far-right column of Table 1.

By accounting for each detected emission line’s wavelength, flux, EW, and, in the case of most [O II] emitters \( (z < 0.42) \), additional lines present in the galaxy spectrum, the overall ability of our Bayesian method to identify high-redshift LAEs targeted by HETDEX for cosmological study is only mildly affected by a mismatch between the expected and actual distributions of galaxy properties. This reflects a practical situation in which we do not know the luminosity functions of the “real” populations with high precision at the onset of the survey.
assuming a typical 0.3 mag reduction in equal-time depth going from \( g' \) to \( r' \). With the Bayesian method, the improvement in changing from \( g' \) (5\( \sigma \) = 25.1) to \( r' \) (5\( \sigma \) = 24.8) band imaging is a ∼1% reduction in \( d_A \). To place this improvement in perspective, switching from \( g' \) to \( r' \) reduces \( d_A \) by ∼1.5% with the EW method, i.e., the observed distributions of LAEs and [O II] emitters are less similar with \( r' \) band imaging. Figure 7 compares simulated \( g' \) and \( r' \) band imaging for a single realization of a simulated spectroscopic survey.

At higher redshifts (2.5 < \( z \) < 3.5), suppressing contamination in \( g' \) band is problematic for \( W_{\lambda_{\text{Ly}\alpha}} > 20 \) Å. The relatively red colors of [O II] emitters (Bridge et al. 2015) lead to weaker continuum flux density measurements in the \( g' \) band, resulting in larger photometric EWs for [O II] determined by aperture photometry on \( g' \) band imaging (Equation (3)). This results in a higher rate of contamination in the LAE sample selected by the traditional EW method versus the case of \( r' \) band imaging. In contrast, our Bayesian method is able to optimize its requirement for LAE classification given the selection of a broadband filter.

4.4. Splitting Available Observing Time to Obtain Color

Inclusion of the colors of galaxies as additional input to the Bayesian method results in modest further reductions in measurement uncertainty in angular diameter distance. In 2.5 < \( z \) < 3.5, splitting available observing time between \( g' \) and \( r' \) in our simulations and using the distributions of \( g' \) – \( r' \) colors of LAEs and [O II] emitters found by HPS (Bridge et al. 2015) as additional Bayesian priors result in ∼5% improvement in \( \sigma_{d_A} / d_A \) over devoting all available observing time to obtaining \( g' \) band imaging in full (25.1 mag) depth.
Tests of other broadband filter combinations yield similar results.

4.5. Other Sources of Contamination

Each of the “additional” emission lines in [O II] spectra targeted for aperture spectroscopy (Section 3.3) may in fact be a 5σ line detection in its own right. However, detections of Hβ λ4861, [O III] λλ4959, and/or [O III] λλ5007 nearly guarantee that [O II] λ3727 will also be detected within the HETDEX spectral range. Therefore, [O III] will not be a significant source of contamination in the HETDEX survey (Adams et al. 2011).

AGNs represent another potential source of confusion for identifying LAEs. In most cases, detection of strong C IV λ1549 provides an indication of the source’s nature; C IV shifts into the HETDEX spectral range at z = 1.25. C III] λ1909 shifts out of the red end of the HETDEX range at z = 1.88, but not before Lyα λ1216 shifts into range at the blue end of HETDEX. Hence when C IV is observed, we always expect to detect C III] or Lyα.

Spurious emission-line detections caused by cosmic rays represent a significant source of contamination. Our Bayesian method can be broadened to account for these spectroscopic detections that have no counterpart in the imaging survey; doing so will greatly increase the importance of deep imaging.

4.6. Application to Future Surveys

4.6.1. Narrowband Surveys

In our application (HETDEX), continuum emission is not well measured in spectroscopy for emission-line-selected objects, necessitating a complementary broadband imaging survey for redshift classification of targeted objects. In addition to spectroscopic emission-line surveys, the Bayesian method presented in this work is equally applicable to narrowband surveys.

There are two limiting cases. In the first, where the narrowband filters have top-hat transmission profiles (e.g., Ciardullo et al. 2012), each object’s line flux will be well determined, while the precision of its redshift measurement is limited to the width of the filter. In the second case, the filter transmission curves may be Gaussian (e.g., Gronwall et al. 2007). In this scenario, there is an additional probability associated with the location of the emission line within the filter bandpass, which creates uncertainty in converting narrowband flux density excess into an EW. In either case, ancillary
broadband data are required to determine the EWs of detected emission lines. Redshift classifications can then be made to identify galaxies targeted by the survey.

4.6.2. Euclid and Wide-Field Infrared Survey Telescope (WFIRST)

Future surveys by the Euclid space mission and the space-based WFIRST will conduct cosmological studies via BAO measurements with slitless spectroscopy for Hα emission-line galaxies at $0.7 < z < 2.1$ (Laureijs et al. 2011) and $1.3 < z < 2.7$ (Green et al. 2012), respectively. Due to their nature in targeting a specific line for detection, these types of investigations are generally susceptible to contamination by other strong line emissions. For example, Geach et al. (2008, 2010) found that [O III] λ5007, Paα, Paβ, and [Fe II] may all be confused with Hα in low-resolution surveys.

A similar Bayesian method can assist these projects, which will have many broadbands available. For application to these upcoming experiments, the method should be broadened to include photometric redshift probabilities in addition to the luminosity functions and EW distributions considered in this study.

When broadband photometric redshifts are available (Pullen et al. 2016), they can be combined with our EW-based probabilities to yield a more robust classification, via $R_{\text{total}} = R_{\text{photo-z}} \times R_{\text{EW}}$, where $R$ refers to posterior odds ratios of LAE versus [O II]. However, these quantities share the information of the continuum flux density near the emission-line wavelength, so they are not completely independent, and hence it would be preferable to perform a joint analysis such as template-fitting including emission-line information.

4.7. Use of Classification Probabilities in Large-scale Structure Analyses

The parameterization of $\sigma_{dA}/dA$ given by Equation (26) implicitly assumes that $dA$ is produced by way of the power spectrum calculated from point estimates of redshift (correspondingly, of object label) based on a cut in classification probability. If we instead consider the catalog to comprise observed objects each with a known classification probability, we may retain potentially valuable information in calculating summary statistics for both populations.

Instead of reducing a probabilistic catalog to a traditional classification problem, we may do hierarchical inference directly on the classification probabilities to obtain the contamination fraction necessary for calculation of the power spectrum (A. I. Malz & D. W. Hogg, 2016, in preparation). An alternative and more ambitious approach would convert the classification probabilities to redshift probabilities, thereby replacing the density field necessary for the calculation of the power spectrum with one that treats each object as having a probability distribution over redshift, a sum of components proportional to the classification probabilities at the redshifts corresponding to the galaxy types in question. Both of these approaches would preserve the knowledge of the classification probabilities but would require greater computational cost, as the calculation of the two-point correlation function must take more information into account.

5. Conclusions

The Bayesian method presented in this work for the classification of LAEs offers robust improvements over the traditional limit requiring LAEs to have rest-frame EW ($W_{\text{Ly}\alpha}$) greater than 20 Å. The statistical discriminating power of our Bayesian method derives from the comoving volumes of the corresponding redshifts based on the assumed cosmology, the properties measured for previously observed samples of LAEs and [O II] emitters, and known positions of other emission lines in the spectra of [O II] emitters. For a simulated HETDEX catalog with realistic measurement noise, our Bayesian method

1. recovers 86% of LAEs missed by the $W_{\text{Ly}\alpha} > 20$ Å cutoff over $2 < z < 3$;
2. outperforms $W_{\text{Ly}\alpha} > 20$ Å in limiting contamination in the LAE sample and increases the completeness of the statistical sample;
3. allows trade off between contamination and incompleteness in arbitrary wavelength/redshift bins.

For a simulated HETDEX catalog, Table 3 demonstrates that our implementation of a Bayesian method reduces uncertainties in angular diameter distance measurements by 14%, which is equivalent to obtaining 29% more data, compared to the $W_{\text{Ly}\alpha} > 20$ Å criterion.

Additional conclusions of our investigation are as follows.

1. For fixed spectroscopic depths, performance of the Bayesian method is relatively insensitive to imaging survey depth, suggesting that maximizing imaging survey area should be favored in a fixed amount of observing time for the purpose of LAE–[O II] galaxy separation.
2. Inclusion of the colors of galaxies as an input to the Bayesian method increases discriminating power and results in modest further reductions in distance errors.
3. Bayesian method can also be used to determine which single broadband filter produces the best performance.
4. Bayesian method can be directly applied to other surveys where single emission lines require classification, including planned space-based observations by Euclid and WFIRST.

Unlike the Bayesian approach, machine-learning methods do not require prior assumptions on the luminosity functions and EW distributions of galaxies, but they require a sizable training set, consisting of ancillary data for which the object labels are known for 5%–10% of the survey (Acquaviva et al. 2014). As a result, we anticipate a combination of these two complementary classification approaches to yield additional improvements.

We thank the referee, Martin Weinberg, for providing critical comments and detailed suggestions, which have led to substantial improvement in the clarity of the method presented in this paper. E.G. and A.S.L. gratefully acknowledge support from the NSF through grant AST-1055919. E.K. is supported in part by MEXT KAKENHI grant No. 15H05896. J.I.F. thanks the University of Texas at Austin for hospitality and support during a sabbatical visit, where a portion of this work was completed.

Appendix

The Effect of Contamination by [O II] Emitters

When [O II] λ3727 lines are misidentified as Lyα λ1216 lines, we erroneously map the correlation function of [O II] galaxies at low redshifts to the correlation function of LAEs at high redshifts. As a result, instead of measuring the correlation at some comoving separations between LAEs at $z > 1.9$, we
actually measure the correlation of [O II] emitters at smaller comoving separations at $z < 0.5$.

This effect can be formulated most straightforwardly using real space correlation functions. Let us imagine that, from the HETDEX data, we measured the correlation function, $\xi(\theta d_A, s_||)$, where $\theta$ is the angular separation on the sky, $d_A$ is the comoving angular diameter distance to a given redshift (assigned to LAEs, including misidentified [O II] emitters), and $s_||$ is the line-of-sight separation.

A problem arises because, since we misidentified the [O II] lines for Ly$\alpha$ lines, our assumed values of $d_A$ and $s_||$ are incorrect.

Let us denote the correlation function of the contamination at $z > 1.9$ as $\xi_{\text{cont.}}$, and the true correlation function of the [O II] emitters at $z < 0.5$ as $\xi_{[O \text{ II}]}$. As these are the same objects, except for the values of the spatial coordinates assigned to each, we have a trivial relation:

$$\xi_{\text{cont.}}(\theta d_A, s_||) = \xi_{[O \text{ II}]}(\theta d_A^\text{[O II]}, s_||^\text{[O II]}).$$

(30)

The angular separation, $\theta$, is the same for both $\xi_{\text{cont.}}$ and $\xi_{[O \text{ II}]}$. The misidentification of the lines simply produces incorrect values for $d_A$ and $s_||$ on the left-hand side.

The power spectrum can be obtained from the correlation function via the usual inverse Fourier transform,

$$P_{\text{cont.}}(k_L, k_0) = \int d^2 s_L e^{-i k_L s_L} \times \int ds_|| e^{-i k_0 s_||} \xi_{\text{cont.}}(s_L, s_||),$$

(31)

where $s_L \equiv \theta d_A$. We also Fourier transform $\xi_{\text{cont.}}(s_L, s_||)$,

$$\xi_{\text{cont.}}(s_L, s_||) = \xi_{[O \text{ II}]}(s_L^\text{[O II]}, s_||^\text{[O II]}) = \int \frac{d^2 k_L^\text{[O II]}}{(2\pi)^2} e^{i k_L^\text{[O II]} s_L^\text{[O II]}} \times \int \frac{dk_0^\text{[O II]}}{2\pi} e^{i k_0^\text{[O II]} s_||^\text{[O II]}} P_{[O \text{ II}]}(k_0^\text{[O II]}).$$

(32)

Now, define the key parameters, the ratios of $s_L$ and $s_||$, as follows:

$$\alpha \equiv \frac{s_L}{s_||^\text{[O II]}} = \frac{d_A}{s_||^\text{[O II]}} > 1,$$

(33)

$$\beta \equiv \frac{s_||}{s_||^\text{[O II]}} = \frac{(1+z)/H(z)}{(1+z_{O \text{ II}})/H(z_{O \text{ II}})} = \frac{3727 H(z_{O \text{ II}})}{1216 H(z)} = \frac{3727}{1216} \frac{\Omega_m (1+z_{O \text{ II}})^3 + \Omega_\Lambda}{\Omega_m (1+z)^3 + \Omega_\Lambda}. $$

(34)

The quantity $\alpha$ is always larger than 1 (i.e., the redshift of LAEs is always higher than that of [O II] emitters), while $\beta$ can be larger or smaller than 1. Of course, $1+z_{O \text{ II}}$ satisfies the relation

$$1+z_{O \text{ II}} = (1+z) \frac{1216}{3727} \approx \frac{1+z}{3.065}. $$

(35)

In Table 5, we summarize the values of $\alpha$ and $\beta$ for representative redshifts $z = 2.2$ and 3.0.

| $z$ | $z_{O \text{ II}}$ | $\alpha$ | $\beta$ |
|-----|-----------------|--------|--------|
| 2.2 | 0.044           | 3900   | 130.7  |
| 3.0 | 0.305           | 4554   | 853.3  |

Notes.

a) In units of $h^{-1}$ Mpc.

b) $\alpha \equiv d_A / s_||^\text{[O II]}$

c) $\beta \equiv (1+z)/H(z)/H(z_{O \text{ II}})$

Using Equation (32) in Equation (31) along with Equations (33) and (34), we find

$$P_{\text{cont.}}(k_L, k_0) = \int d^2 k_L^\text{[O II]} \int d^2 s_L e^{i k_L^\text{[O II]} s_L} \int \frac{dk_0^\text{[O II]}}{2\pi} P_{[O \text{ II}]}(k_0^\text{[O II]}).$$

(36)

This is the equation we use for computing the contamination of the LAE power spectrum. This result makes physical sense for the following situations.

1. For a given set of values of $k_L$ and $k_0$, we are actually observing the power spectrum of [O II] emitters at smaller scales, $k_L \rightarrow \alpha k_L$ and $k_0 \rightarrow \beta k_0$. This contamination produces a horizontal shift of the [O II] power spectrum to smaller $k$ values.

2. As the correlation function measures the dimensionless power spectrum, $k^2 P(k)$, the normalization of the power spectrum is also shifted by $\alpha^2 \beta$, generating a vertical shift of the [O II] power spectrum.

At lower redshifts, $\alpha$ can be large as $\alpha \approx 30$ (see Table 5), which boosts the amplitude of the [O II] power spectrum by a factor of $\alpha^2 = 900$. Conversely, $\beta$ is on the order of unity for the redshift range of interest.

Finally, the observed power spectrum is given by the weighted average of the LAE power spectrum, $P_{\text{LAE}}$, and the contamination power spectrum:

$$P_{\text{obs}}(k_L, k_0) = (1 - f_{O \text{ II}})^2 P_{\text{LAE}}(k_L^2 + k_0^2) + f_{O \text{ II}}^2 (\alpha^2 \beta) P_{[O \text{ II}]}(k_L^2 + \beta^2 k_0^2).$$

(37)

where $f_{O \text{ II}}$ is the fraction of [O II] emitters in the total sample, i.e.,

$$f_{O \text{ II}} \equiv \frac{\text{number of contaminating [O II] emitters}}{\text{number of galaxies classified as LAEs}}.$$
References

Acquaviva, V., Gawiser, E., Leung, A. S., & Martin, M. R. 2014, in IAU Symp. 306, Statistical Challenges in 21st Century Cosmology, ed. A. Heavens, J.-L. Starck, & A. Krone-Martins (Lisbon: Portugal: IAU), 365

Adams, J. J., Blanc, G. A., Hill, G. J., et al. 2011, ApJS, 192, 5

Anders, P., & Fritze-v. Alvensleben, U. 2003, A&A, 401, 1063

Blanc, G. A., Adams, J. J., Gebhardt, K., et al. 2011, ApJ, 736, 31

Bridge, J. S., Gronwall, C., Ciardullo, R., et al. 2015, ApJ, 799, 205

Ciardullo, R., Gronwall, C., Adams, J. J., et al. 2013, ApJ, 769, 83

Gawiser, E., van Dokkum, P. G., Gronwall, C., et al. 2006, ApJL, 642, L13

Geech, J. E., Cimatti, A., Percival, W., et al. 2010, MNRAS, 402, 1330

Geach, J. E., Smail, I., Best, P. N., et al. 2008, MNRAS, 388, 1473

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 1995, Bayesian Data Analysis (Boca Raton, FL: CRC Press)

Green, J., Schechter, P., Balty, C., et al. 2012, arXiv:1208.4012

Gronwall, C., Ciardullo, R., Hickey, T., et al. 2007, ApJ, 667, 79

Gualtieri, L., Gawiser, E., Padilla, N., et al. 2010, ApJ, 714, 255

Henry, A. L., Martin, C. L., Dressler, A., Sawicki, M., & McCarthy, P. 2012, ApJ, 744, 149

Hill, G. J., & HETDEX Consortium 2016, in ASP Conf. Ser. 507, Multi-Object Spectroscopy in the Next Decade: Big Questions, Large Surveys, and Wide Fields ed. I. Skillen, M. Barcels, & S. C. Trager (San Francisco, CA: ASP), 393

Hill, G. J., Gebhardt, K., Komatsu, E., et al. 2008, in ASP Conf. Ser. 399, Panoramic Views of Galaxy Formation and Evolution, ed. T. Kodama, T. Yamaeda, & K. Aoki (Hayama, Japan: ASP), 115

Hippelein, H., Maier, C., Meisenheimer, K., et al. 2003, A&A, 402, 65

Hogg, D. W., Cohen, J. G., Blandford, R., & Pahre, M. A. 1998, ApJ, 504, 622

Hu, W., & Haiman, Z. 2003, PhRvD, 68, 063004

Koehler, R. S., Schuecker, P., & Gebhardt, K. 2007, A&A, 462, 7

Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, ApJS, 192, 18

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193

Ly, C., Malkan, M. A., Kashikawa, N., et al. 2007, ApJ, 657, 738

Lynd, R. 1971, ApJL, 164, L73

Madau, P. 1995, ApJ, 441, 18

Oke, J. B., & Gunn, J. E. 1983, ApJ, 266, 713

Osterbrock, D. E. 1974, Astrophysics of Gaseous Nebulae (San Francisco: Freeman)

Ouchi, M., Shimasaku, K., Akiyama, M., et al. 2008, ApJS, 176, 301

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, 13

Pullen, A. R., Hirata, C. M., Doré, O., & Raccanelli, A. 2016, PASI, 68, 12

Schechter, P. 1976, ApJ, 203, 297

Seo, H.-J., & Eisenstein, D. J. 2003, ApJ, 598, 720

Seo, H.-J., & Eisenstein, D. J. 2007, ApJ, 665, 14

Shimasaku, K., Kashikawa, N., Doi, M., et al. 2006, PASJ, 58, 313

Shoji, M., Iwamuro, D., & Komatsuda, T. 2009, ApJ, 693, 1404

Takahashi, M. I., Shioya, Y., Taniguchi, Y., et al. 2007, ApJS, 172, 456

Teplitz, H. I., Collins, N. R., Gardner, J. P., Hill, R. S., & Rhodes, J. 2003, ApJ, 589, 704

Yan, R., Newman, J. A., Faber, S. M., et al. 2006, ApJ, 648, 281