The Astrophysical Journal, 890:74 (17pp), 2020 February 10 https://doi.org/10.3847/1538-4357/ab684f

Applying Information Theory to Design Optimal Filters for Photometric Redshifts

J. Bryce Kalmbach[1], Jacob T. VanderPlas[2], and Andrew J. Connolly[1]
[1] DIRAC Institute and Department of Astronomy, University of Washington, Seattle, WA 98195, USA; brycek@uw.edu
[2] The eScience Institute, University of Washington, Seattle, WA 98195, USA
Received 2019 July 12; revised 2020 January 3; accepted 2020 January 5; published 2020 February 13

Abstract

In this paper we apply ideas from information theory to create a method for the design of optimal filters for photometric redshift estimation. We show the method applied to a series of simple example filters in order to motivate an intuition for how photometric redshift estimators respond to the properties of photometric passbands. We then design a realistic set of six filters covering optical wavelengths that optimize photometric redshifts for $z <= 2.3$ and $i < 25.3$. We create a simulated catalog for these optimal filters and use our filters with a photometric redshift estimation code to show that we can improve the standard deviation of the photometric redshift error by 7.1% overall and improve outliers 9.9% over the standard filters proposed for the Large Synoptic Survey Telescope (LSST). We compare features of our optimal filters to the LSST and find that the LSST filters incorporate key features for optimal photometric redshift estimation. Finally, we describe how information theory can be applied to a range of optimization problems in astronomy.

Unified Astronomy Thesaurus concepts: Astrostatistics (1882); Galaxy photometry (611); Galaxy distances (590); Astronomy data analysis (1858); Surveys (1671)

1. Introduction

In a seminal work, Shannon (1948) introduced the concept of information theory. While originally concerned with the information content of messages sent along a channel with limited bandwidth and other signal processing problems, applications of information theory now extend to a multitude of fields, including finance (Ormos & Zibriczky 2014) and genomics (Adami 2004). Information theoretic concepts are now used in astronomy across a wide range of problems. For instance, Weir et al. (1995) worked with decision trees for star/galaxy classification and used the information entropy to inform the class impurity at each branching. In Seehars et al. (2014) the authors utilized information theory to judge the information gain (IG) on parameter posteriors from a series of cosmic microwave background experiments. They were also able to separate the information gained by improvements in statistical error from that gained by new data changing the posterior distributions. Cincotta et al. (1995) proposed the use of Shannon entropy to find the period of astronomical light curves, and Graham et al. (2013b) extended this to use conditional entropy. Graham et al. (2013a) showed that the conditional entropy algorithm was the best compared to a wide variety of other period-finding methods, including Lomb–Scargle with regard to period recovery and computation time. Finally, Huijse et al. (2018) used mutual information to combine light curves measured in different photometric bands and recover periods more effectively than multiband extensions of Lomb–Scargle and analysis-of-variance periodograms. In this paper, we apply information theory to a combination of survey design and photometric redshift estimation problems.

Photometric redshift estimation uses multiple observations of extragalactic sources, spread across a range of filters or passbands, to derive an approximate redshift for a given source (Baum 1962; Koo 1985; Connolly et al. 1995). The upcoming Large Synoptic Survey Telescope (LSST; Ivezić et al. 2019) will rely on photometric redshifts for the vast majority of galaxies imaged in the course of its 10 yr survey. The accuracy of these redshift estimates is dependent on the position of breaks or features within a source spectrum relative to the passbands of the photometric filters. For example, the 4000 Å break transitions out of the LSST y band at a redshift of $z \sim 1.5$, resulting in an increase in the uncertainty of the estimated redshifts until the Lyman break enters the u band at $z \sim 2.5$. In principle, the location and shape of a set of filters could be designed to track specific features within a galaxy spectrum and thereby improve the photometric redshift (at least over a narrow range of redshifts). This work attempts to find a principled way to define the photometric redshift performance of optical filters using information theory and, more specifically, IG and thereby derive a set of filters that are optimal for a specific set of survey objectives. The IG method we outline here can be extended in the future to other areas of astronomy where color can be used to classify objects. Here our classes are photometric redshift bins but could easily be used to classify types of stars instead.

We start in Section 2 with a brief primer on information theory before applying the concept to three basic examples in the following sections. In Section 6 we describe the algorithm and code we developed to calculate IG for astronomical filters. In Section 7 we apply the technique to design optimal filter sets, and in Section 8 we compare photometric redshifts for a simulated catalog using the proposed filter sets versus LSST filters. We discuss our work and future directions for it in Section 9 and conclude in Section 10.

2. Introduction to Information Theory

2.1. Entropy

Consider an event Y with a set of n possible outcomes y_1, y_2, y_3, …, y_n that each occur with probabilities p_1, p_2, p_3, …, p_n and $\sum_{i=1}^{n} p(y_i) = 1$. How can we measure the amount of choice or uncertainty present in the selection of an outcome? Shannon (1948) concluded that the uncertainty in the observed outcome is given by the entropy (H) of this set of possible outcomes,

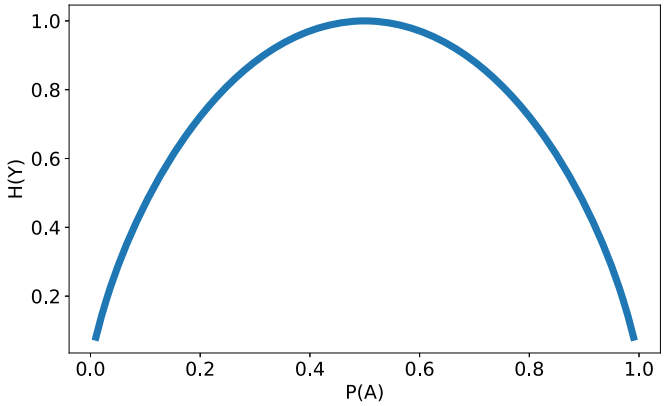

Figure 1. Entropy of a system with two outcomes A and B as the probability of getting outcome A changes.

where the entropy is defined as

$$H(y) = -\sum_i p(y_i)\log_2[p(y_i)]. \quad (1)$$

Some properties that become apparent from Equation (1) are that the maximum entropy occurs when all outcomes are equally probable and that entropy becomes zero when a single probability dominates. Figure 1 shows the entropy when we have two possible choices A and B and how the entropy changes as the probability of getting outcome A changes. When using base 2 in the logarithm, the entropy is measured in bits and represents the average number of binary digits required to encode a set of outcomes from Y.

For instance, imagine we are observing an event that has two possible outcomes that we label A and B. If $p(A) = p(B) = 0.5$, then the entropy calculation tells us that the best representation we can derive will encode $H(A) + H(B) = -2 * 0.5\log_2(0.5) = 1$ bit on average. Therefore, simply using $A = 0$ and $B = 1$ when reporting a string of results is an optimal encoding since there is a one-to-one relationship with the length of the encoded information and the number of results. However, if we had a situation where $p(A) > p(B)$, we would have an entropy less than 1. According to information theory then, the best encoding scheme could encode the results with less than 1 bit on average. To say this in the reverse way, this means that we can represent a string of results with a number of bits smaller than the length of the results string. Unfortunately, knowing the amount of information in the distribution does not tell us how to optimally encode information, but a possible method would be to encode sequences of results with sequences of bits. This would mean that each bit of information on average would represent more than one result.

2.2. Conditional Entropy and Information Gain

Lindley (1956) was the first to extend information theory to quantify the information gained from a measurement by measuring how much an experiment reduced entropy. For instance, imagine that a community wants to screen its members for an illness and we know it targets primarily individuals over 40. If we only have a list of the members of the community, we can only assign the same probability of illness to all members and can do no better than randomly reaching out to individuals in the population. But if we know the ages of the population, we have more information about whom we should target. Using information theory, we can actually measure the information gained when the additional information, in this case the ages of the population, is known. To do so, we need to know that conditional entropy is the amount of entropy in an observation of Y when the value of $X = x_j$ is a known quantity. It is defined mathematically in a similar way to entropy:

$$H(Y|X = x_j) = -\sum_i P(y_i|x_j)\log_2[p(y_i|x_j)]. \quad (2)$$

In our example, Y is the probability of illness in the overall community and X is the age. If we know the overall distribution of X, we can calculate the average conditional entropy for the system:

$$H(Y|X) \equiv \sum_{i,j} P(x_j)H(y_i|x_j). \quad (3)$$

For the example presented here, we have a different probability for an illness at different ages, and this gives us additional information that refines the probability of illness to be more precise for each individual. Therefore, the average conditional entropy will be smaller than the overall entropy since we have less uncertainty in estimates of who might be ill. The actual IG can be found by subtracting the average conditional entropy from the original entropy:

$$\mathrm{IG}(Y|X) = H(Y) - H(Y|X). \quad (4)$$

To put numbers into our example, let us give the overall probability of the illness to be 22%, but for the 60% of the population under 40 the probability becomes only 10%, while for the other 40% of the population it is 40%. Without information on the ages, we then have .76 bits of entropy in our estimates of illness $(.22 * \log_2(.22) + .78 * \log_2(.78) = .76)$. Adding in the age information gives us an average conditional entropy of $(.6 * (.1 * \log_2(.1) + .9 * \log_2(.9))) + (.4 * (.4 * \log_2(.4) + .6 * \log_2(.6))) = .67$ bits. Therefore, the IG becomes $.76 - .67 = .09$ bits of information gained when we incorporate age information. In the extreme that an illness hit everyone over 40 and nobody under 40, then age information would provide us with a perfect understanding of who had the disease and who did not. In this case, average conditional entropy would fall to 0 and we would have IG equal to the total original entropy. This shows that the more IG we can derive from a measurement of X, the more this measurement reduces our uncertainty in another property Y.

2.3. Application to Astronomical Observation

Often in astronomy, we employ a particular observation (be it photometric, spectroscopic, or other) in order to learn about particular properties of the object we are observing. In the formalism expressed above, our observation (say, the magnitude through a particular photometric filter) is given by X, where X represents a continuous distribution of observed values. The intended classification of the object (be it star/quasar classification, galaxy type, photometric redshift, etc.) is represented by the values Y, which may or may not include prior probabilistic information. Given a suitably realistic spectroscopic model of our sample, we can calculate the IG expected from a given filter set and use this quantitative measure to optimize our choice of filters for the task. In the following sections, we will explore the properties of

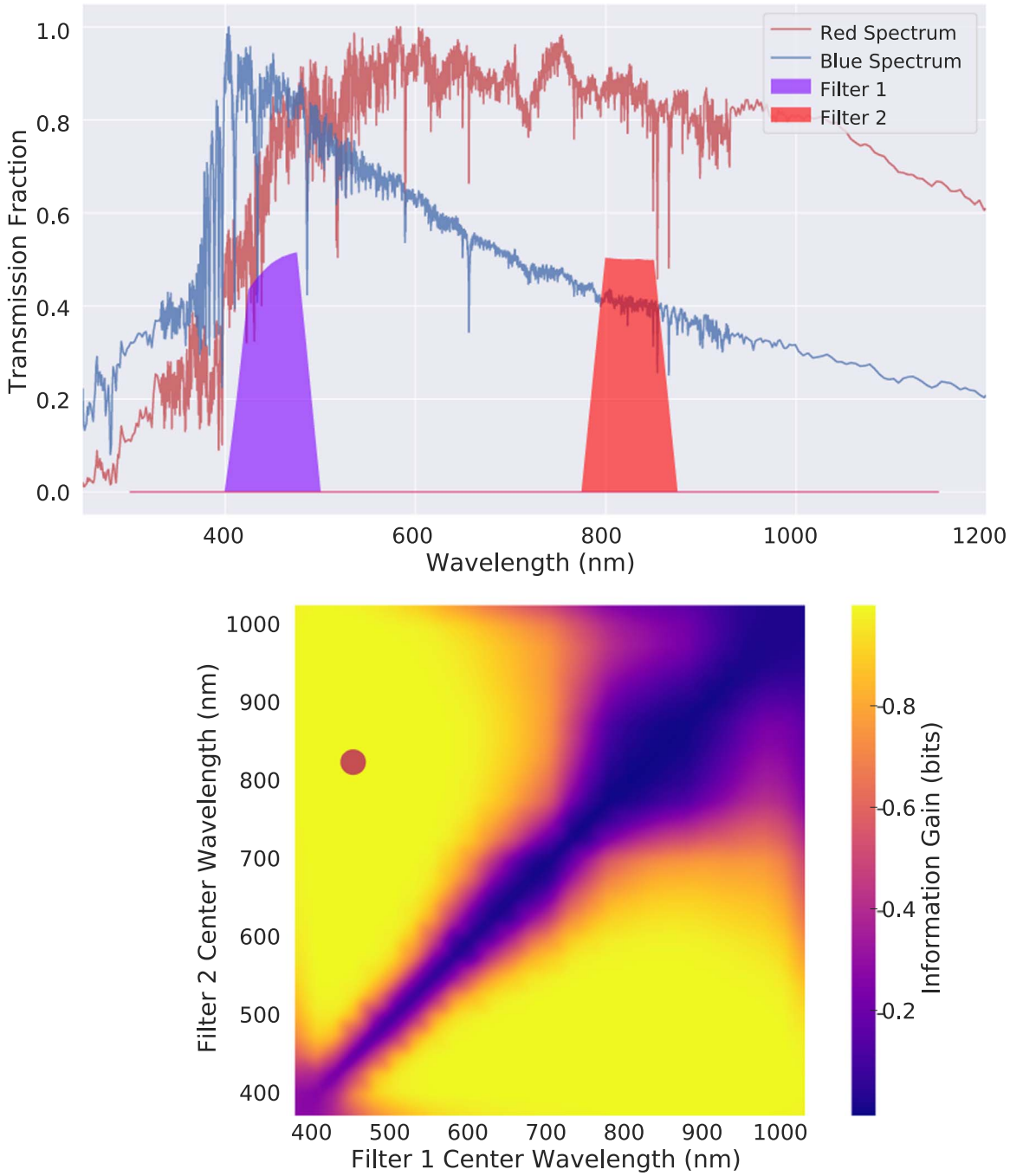

Figure 2. Top: optimal filters for differentiating between equally probable red and blue spectra. Bottom: IG as a function of central wavelength of each filter. Notice that when the filters are nearly identical the IG tends toward 0. The maximum IG filters shown in the top panel are located at the red point with an IG of over 0.99 bits.

IG as applied to increasingly more realistic astronomical measurements.

3. Toy Example 1: Simple Galaxy Classification

Imagine for the time being that all galaxies have spectral energy distributions (SEDs) that fall precisely in one of two classes: we will call them "red" and "blue" (see Figure 2, top panel). We will denote this spectral type by the label Y, which can take on the values $Y \in \{y_r, y_b\}$. Furthermore, imagine that any galaxy has an even chance of being either red or blue. Mathematically stated, this means that $P(y_r) = P(y_b) = 0.5$. From Equation (1) we can quickly compute the entropy $H(Y) = 1$.

Now suppose that an astronomer would like to choose a pair of filters, the magnitude difference (i.e., color) of which will give maximal discrimination between the two types of galaxies. Heuristically, it is clear that placement of one filter toward the left and another toward the right accomplishes this: the difference between the filter magnitudes gives a positive (red) color for spectrum y_r and a negative (blue) color for spectrum y_b, leading to an ability to easily distinguish between the two spectra.

This conclusion can be reached in a quantitative fashion by computing the IG for color measurements through the two filters at various locations as shown in the bottom panel of Figure 2. To construct this surface, we assume trapezoidal

filters (see the top panel of Figure 2) with a total filter response width of 100 nm, containing sloped wings of width 25 nm, and numerically compute magnitudes through each filter. We also include a realistic CCD response function that accounts for the curved edges noticeable in the blue end filter and the very top of the red filter. We normalize the spectra to $i = 22.0$ and include a sky background normalized at $i = 20.47$. Finally, we assume a single LSST visit to calculate the magnitudes and the signal-to-noise ratio (S/N) of each filter. Subtracting the two magnitude values for each spectrum gives us the respective colors with a defined set of filters. We use the S/N to calculate the expected uncertainty around each color measurement. This will give us a Gaussian distribution for the colors of each spectrum in a given filter set. The conditional entropy is then calculated by measuring how much the two color distributions overlap. What we hope to see is that the filter locations that maximize IG are those that move the two color distributions as far apart as possible.

When we look at Figure 2, we see exactly that. The top panel shows that the maximal IG filters are located with peaks around 450 and 825 nm. The bottom panel is a plot of the IG as a function of the center wavelength of each filter and displays that we can almost perfectly distinguish one spectrum from the other since our IG is greater than 0.99 bits out of a possible 1.0. The nearly zero IG along the diagonal makes sense since this is where the filters lie on top of one another and produce the same measured magnitude on average. Information gain is near but not completely zero along this axis since the width of the error distribution in the color measurement is different for each spectrum. Therefore, we do have a little bit of information to help label an observation. For instance, if we make an observation with identical filters and get a color value of 0.02 mag, and this turns out to be a 3σ measurement for the red spectrum but 5σ for the blue spectrum, this provides a small amount of information that increases the probability of this being a red spectrum measurement.

The sharp rise in IG from the diagonal in Figure 2 is a result of the bright galaxies we used. If the galaxies are fainter, then the increased noise means that the red filter must be farther from the peak of the blue spectrum to avoid the possibility of measuring similar colors for both galaxies. This would present itself as a shallower slope in the IG space.

This case showed the basics of the IG theory with an easy problem. Discriminating between two spectra is something we can easily do without resorting to IG, but finding the filters that help discriminate between galaxies at different redshifts is a more realistic and interesting problem. In the next two sections we move on to two simple examples of optimizing filters for photometric redshift estimation.

4. Toy Example 2: Measuring the Redshift of a Sigmoid Spectrum

We can use the same formalism as above to address the question of filter choice for the determination of photometric redshifts. In this case the observable Y is the redshift of the galaxy. Because Y cannot represent a continuous distribution within the IG framework (note the sums in Equation (2) above and see Section 6 below for more information), we must bin the result. In practice, this is not a problem: using a sufficiently large number of bins will allow the redshift result to be recovered to any reasonable accuracy.

For the sake of descriptive simplicity, we will begin with a toy model using very simple spectra. Imagine now that every galaxy in the universe has a spectrum given by a sigmoid function:

$$S(\lambda;\, \lambda_0) = \frac{1}{1 + \exp(\lambda - \lambda_0)}. \tag{5}$$

This is very close to a step function with $S(\lambda) = 0$ for $\lambda \ll \lambda_0$ and $S(\lambda) = 1$ for $\lambda \gg \lambda_0$. With $\lambda_0 = 364.6$ nm, this shape mimics the Balmer-limit break observed in the spectra of galaxies, from which photometric redshift determination gains significant leverage. Imagine furthermore that these galaxies are located at various redshifts, with a probability distribution given by

$$P(z) \propto z^2 \exp[-(z/z_0)^2]. \tag{6}$$

with z_0 set so that the median z is 0.6 (typical of ground-based surveys such as DES; Pogosian et al. 2005). If we break the redshift range into 40 bins from $0.0 < z \leqslant 2.0$ (giving a bin width $\Delta z = 0.05$), then the information contained in the redshift of a galaxy can be computed to be $H(Z) \approx 4.4$ using Equation (1) and the prior. This can be interpreted as saying that, on average, 4.4 bits of information are needed to specify the redshift of a particular galaxy in the distribution. Because there are $40 \approx 2^{5.3}$ bins, one might wonder why a full 5.3 bits per galaxy would not be needed to specify the redshift; the reason for this is due to the prior information contained within the probability distribution (Equation (6)), which allows a more compact representation of the data.

If we perform a maximization of the IG (Equation (4)) using the color observed through two filters as in Section 3, we obtain the IG surface shown in the bottom panel of Figure 3. The locations of the optimal filters are much more constrained than in the binary choice example in Section 3. Because the redshift distribution peaks near $z = 0.55$, filter combinations where the leftmost filter is centered near 600.0 nm lead to the greatest IG, as seen in the top panel of Figure 3. Since the majority of galaxies are located near the peak at $z = 0.55$, a filter that can trace the passage of the Balmer break through redshifts near the peak is the most beneficial. The broadness of the region of maximal IG shows that there is a large region of the parameter space in which the filter locations lead to nearly maximal information. As long as one filter is located to measure the spectral break near the peak of the redshift prior, then the other filter can be shifted left or right over a range of >200 nm without significantly reducing the IG.

Quantitatively, the maximal IG using two filters is ~1.95, out of a total information of roughly 4.40. That is, in this simple model, photometric redshifts based on a single color can recover 44% of the redshift information. Most of the lost information exists because we cannot easily differentiate between redshifts close to one another when the break is outside the filters.

5. Toy Example 3: Measuring the Redshift of a Single Galaxy

Though the sigmoid spectrum explored in Section 4 gives some interesting insight, realistic spectra have many more features in addition to the Balmer break. In this section, we

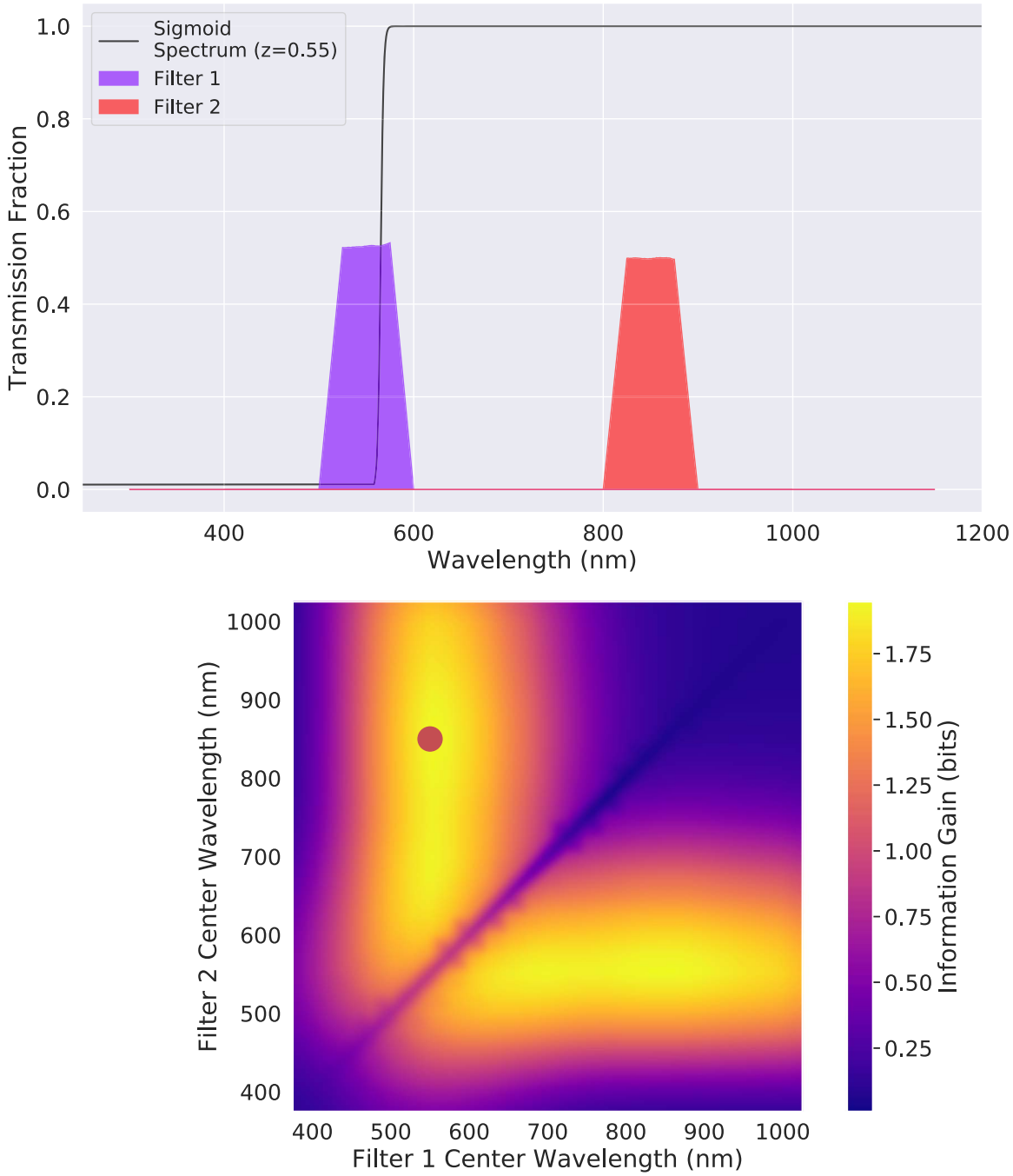

Figure 3. Top: optimal filters for differentiating between the sigmoid spectrum at different redshifts up to $z = 2$. The sigmoid spectrum is shown at a redshift of 0.55 near the peak of the redshift prior function. Bottom: IG as a function of central wavelength of each filter. The maximum IG filters shown in the top panel are located at the red point with an IG of ∼1.95 bits out of a possible 4.4.

explore a similar example using a single red synthetic galaxy spectrum.

Figure 4 shows the equivalent of Figure 3 for this more realistic spectrum. The spectrum is that of the red galaxy from Section 3 with a strong Balmer break around 400 nm. If the redshift information is coming primarily from this break, we would expect the optimal filter locations and associated IG to be similar to that seen in the sigmoid spectrum of Section 4.

As before, the IG surface in Figure 4 shows a region of low IG in the places where the two filters largely overlap. Also like the previous example, the Balmer break is the main factor that determines the locations of the optimal filters. The bluer filter is once again located in the range where the break is passing through the filter when the spectrum is redshifted to the peak redshift of the prior distribution. The maximal IG in this case is slightly higher than what we saw previously: ∼2.19 out of 4.40. This increase is mainly due to the broadness of the Balmer jump (about 100 nm wide here compared to a negligible width previously), which allows a wider range of redshifts to benefit from the change in magnitude of the blue filter as the break passes through it.

To see exactly what is the difference that leads to better IG in one set of filters versus another, we explain the results shown in Figure 5. Here the top panel shows the probability distributions of observed colors of the spectrum at each redshift using the

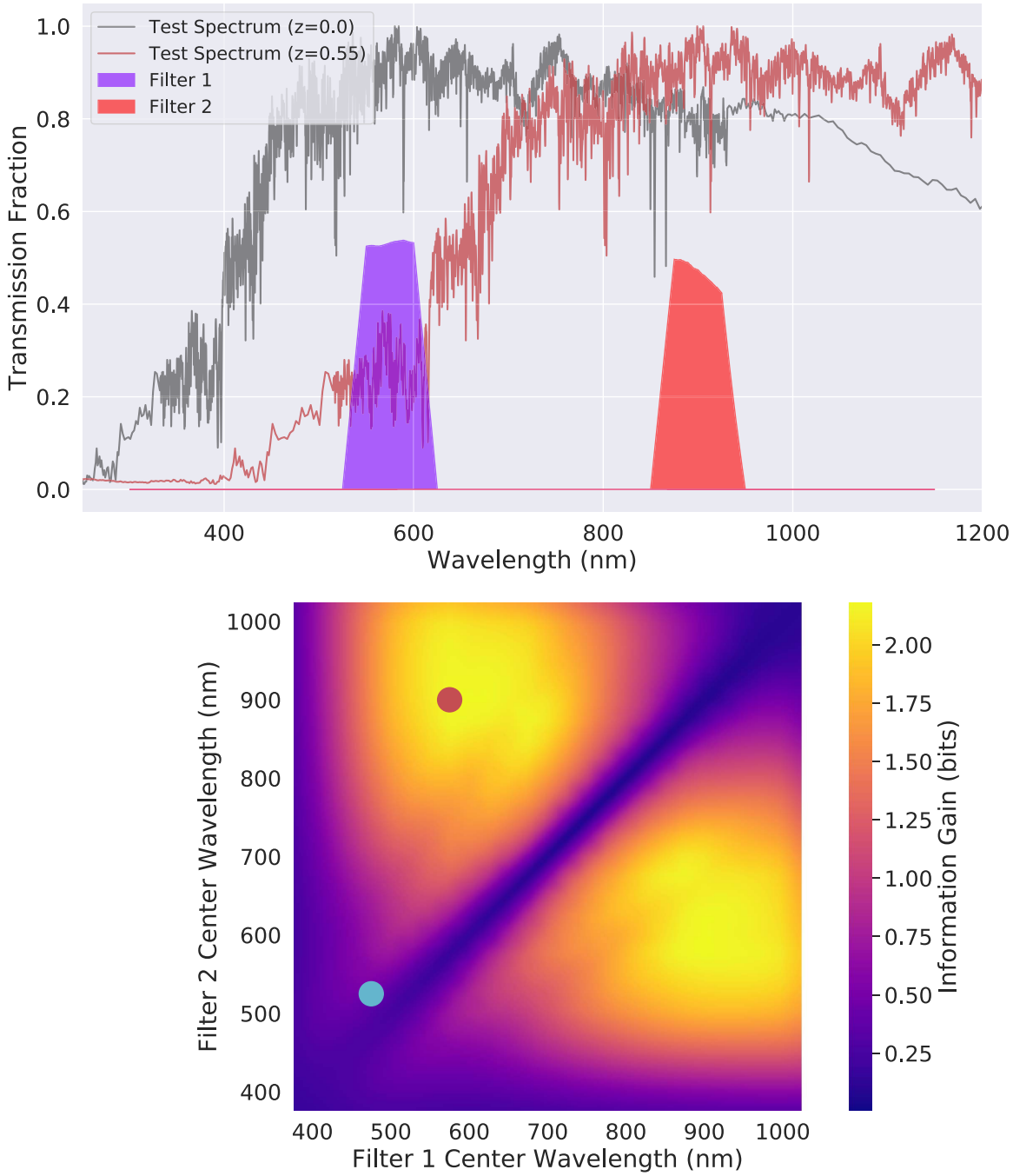

Figure 4. Top: optimal filters for differentiating between the red galaxy spectrum at different redshifts up to $z = 2.5$. Bottom: IG as a function of central wavelength of each filter. The maximum IG filters shown in the top panel are located at the red point with an IG of ∼2.19 bits out of a possible 4.4. The blue point shows the location of the alternate set of filters used in Figure 5.

optimal filters from Figure 4 weighted by the prior probability at that redshift. These probability distributions are centered at the mean color for the template at a given redshift. Since the redshift is binned and calculated in steps of 0.05, we see a discrete set of color distributions. The width of each color distribution is a result of photometric uncertainties and is affected by the design of the filters and survey. In the bottom panel we see the same distributions for colors derived using a set of filters that only produced 0.46 bits of IG, and they are marked in the bottom panel of Figure 4 by the blue circle. Notice how much more overlap there is in the distributions for colors at each redshift in the bottom panel. On top, where we have higher IG, we can be much more confident that a galaxy measured with a certain color will have a given redshift, especially in the redshifts around the peak of the prior distribution at $z \sim 0.55$.

The simple examples shown here help connect the information theory presented to practical results in astronomical terms. However, to fully apply information theory to larger template sets and higher numbers of filters, we need to further develop our mathematical approach and build the computational tools that will allow us to perform larger experiments.

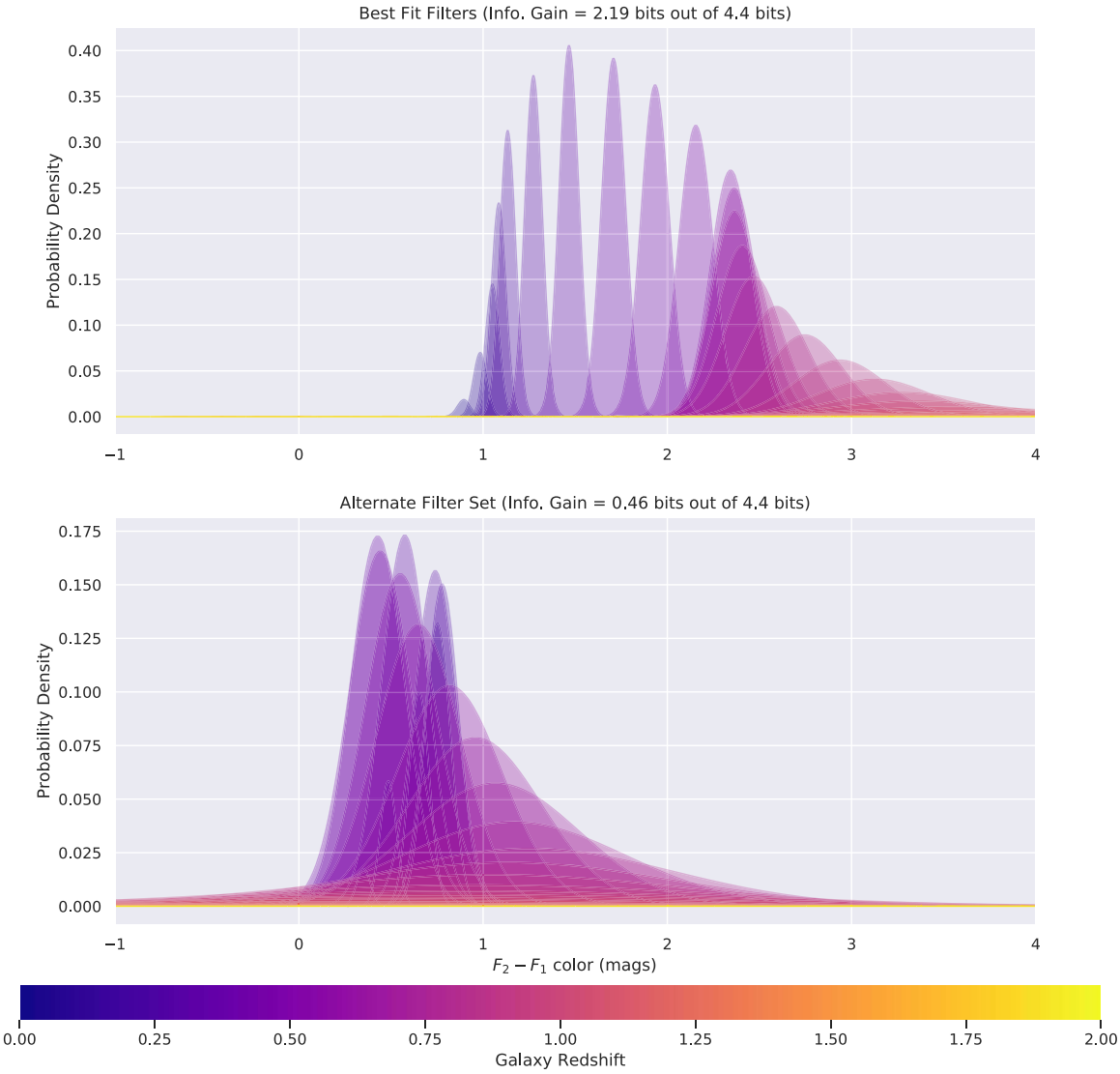

Figure 5. Top: distribution of colors for the red galaxy spectrum at a series of redshifts using the optimal filter scheme. This figure includes the prior on redshift that more strongly weights intermediate redshifts over low and high redshifts. Notice how redshifts near the peak of the prior ($z \sim 0.55$) have the least overlap in their possible color values. Bottom: distribution of colors using a filter scheme that produces 21% of the optimal IG (this corresponds to the blue point in Figure 4). Here the distributions pile on top of one another and a given color could be the result of the spectrum at a large number of possible redshifts.

6. Calculating Information Gain in Practice

To calculate the IG in more complicated scenarios, we needed to develop code that could quickly calculate IG based on multiple colors and redshifts of multiple SEDs. For this purpose we created a python code called SIGgi (where SIG stands for spectral information gain; Kalmbach 2019). In practice, we calculate IG starting from calculating the average conditional entropy $H(Y|X)$, where we rewrite it by combining Equations (2) and (3) along with the identity $P(x_i, y_i) = P(y_i|x_i)P(x_i)$ to get

$$H(Y|X) = -\sum_{i,j} p(x_j, y_i)\log_2\left[\frac{p(x_j, y_i)}{p(x_j)}\right]. \quad (7)$$

But in our case X is the vector of colors of the SED and is continuous. Therefore, we allow continuous observations by using the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951):

$$D_{\mathrm{KL}}(P||Q) = \sum_k p(y_k)\log_2(p(y_k)/q(y_k)). \quad (8)$$

If $p(y)$ and $q(y)$ are continuous probability distributions and normalized to 1 across the entirety of y, then the KL divergence is

$$D_{\mathrm{KL}}(p||q) = \int p(y)\log_2[p(y)/q(y)]\, dy. \quad (9)$$

Now we can see that $H(Y|X)$ can be written in terms of the KL divergence as

$$H(Y|X) = -D_{\mathrm{KL}}(p(y_i, x)||p(x)), \quad (10)$$

where y remains a discrete variable and thus requires that we continue to bin redshift, but now x is expressed as a continuous

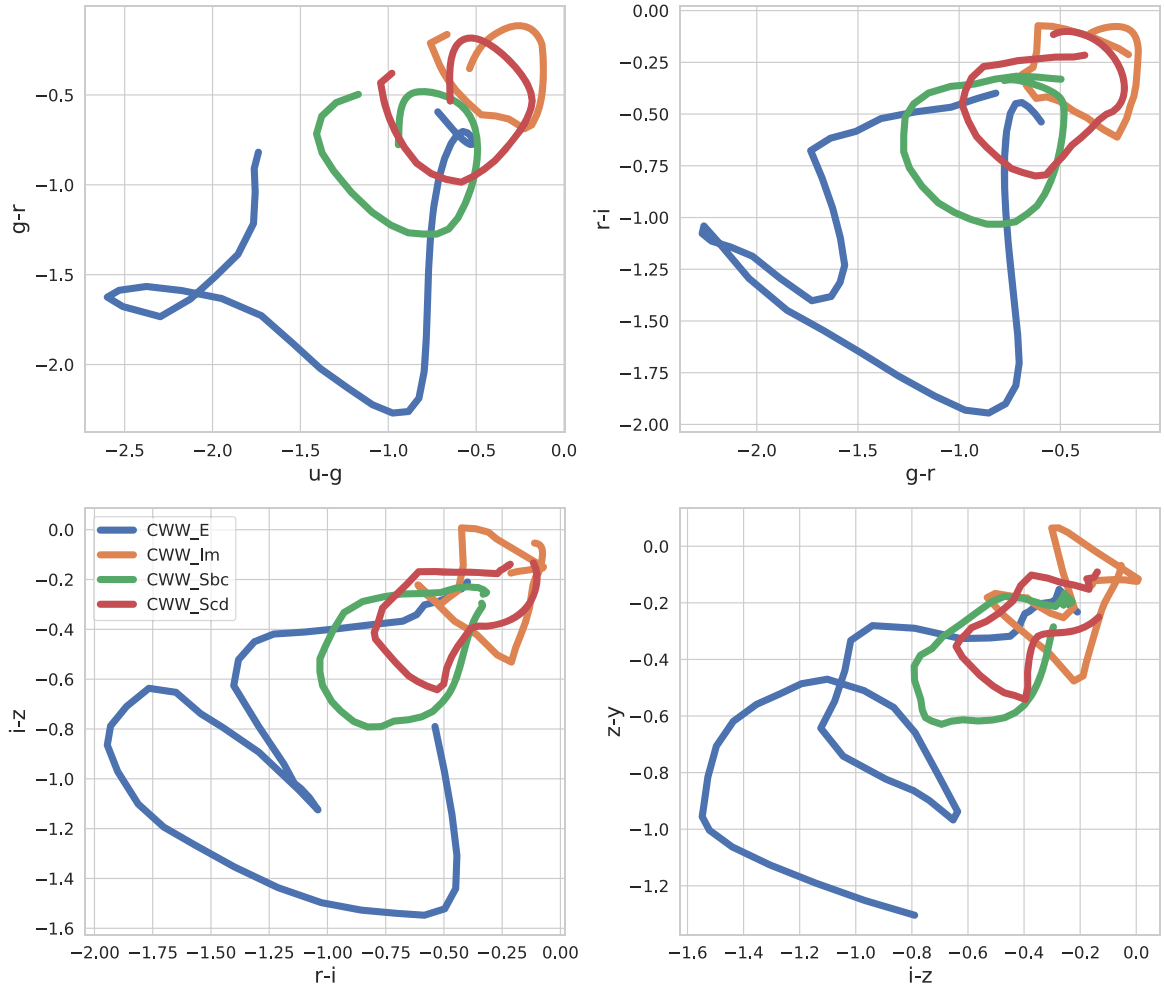

Figure 6. Color–color plot of the four Coleman et al. (1980) templates in the LSST filters.

observable. Finally, we combine this with Equation (7) to get

$$H(Y|X) = -\sum_i \int p(y_i, x) \log_2 \left[\frac{p(y_i, x)}{p(x)} \right] dx, \qquad (11)$$

where i is a particular redshift bin.

So, to compute the conditional entropy, we must be able to determine the joint probability $P(y_i, x) = P(y_i)P(x|y_i)$, where each y_i represents a discrete unknown property (e.g., binned redshift) and x is a continuous observable (e.g., photometric colors). $P(y_i)$ is simply the prior distribution of the unknown property, and $P(x|y_i)$ expresses the distribution of observables for a particular input. We have a model to predict this conditional distribution $P(x|y_i)$ of an observation x given a value y_i (e.g., we can compute the colors of a galaxy given its redshift). In the case of a single spectrum this distribution $P(x|y_i)$ is assumed to be normally distributed about a single value. The width of the Gaussian in each dimension will depend on the photometric uncertainty of the measurements in the filters for that color. In the case of multiple spectra the color distribution for a single redshift will be the sum of the normal distributions for each individual spectrum at that redshift.

Because the calculation of conditional entropy via Equation (11) involves an integral over a potentially high-dimensional space with very fine resolution, a straightforward numerical integration based on a grid of values becomes too costly to use in practice. For example, a single color for a collection of galaxy spectra in LSST filters may spread over up to two and a half magnitudes (see Figure 6). To assure sufficient sampling of the distribution of colors, this requires on order of 100 grid divisions per dimension, which leads to on order of 10^{10} grid points in five dimensions for a naive

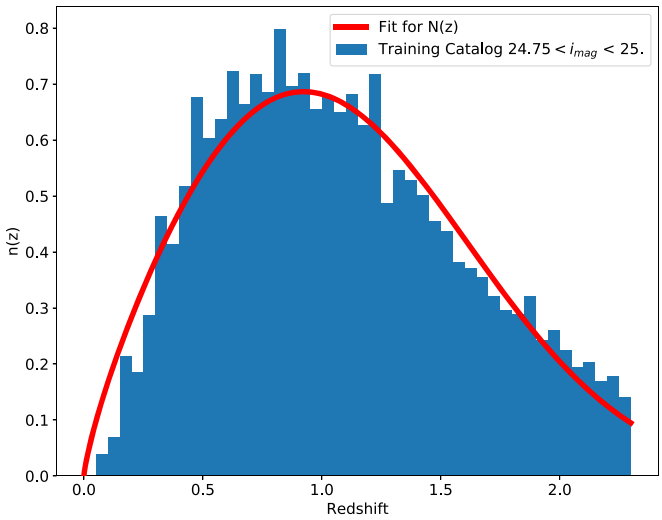

Figure 7. Redshift prior derived from training catalog.

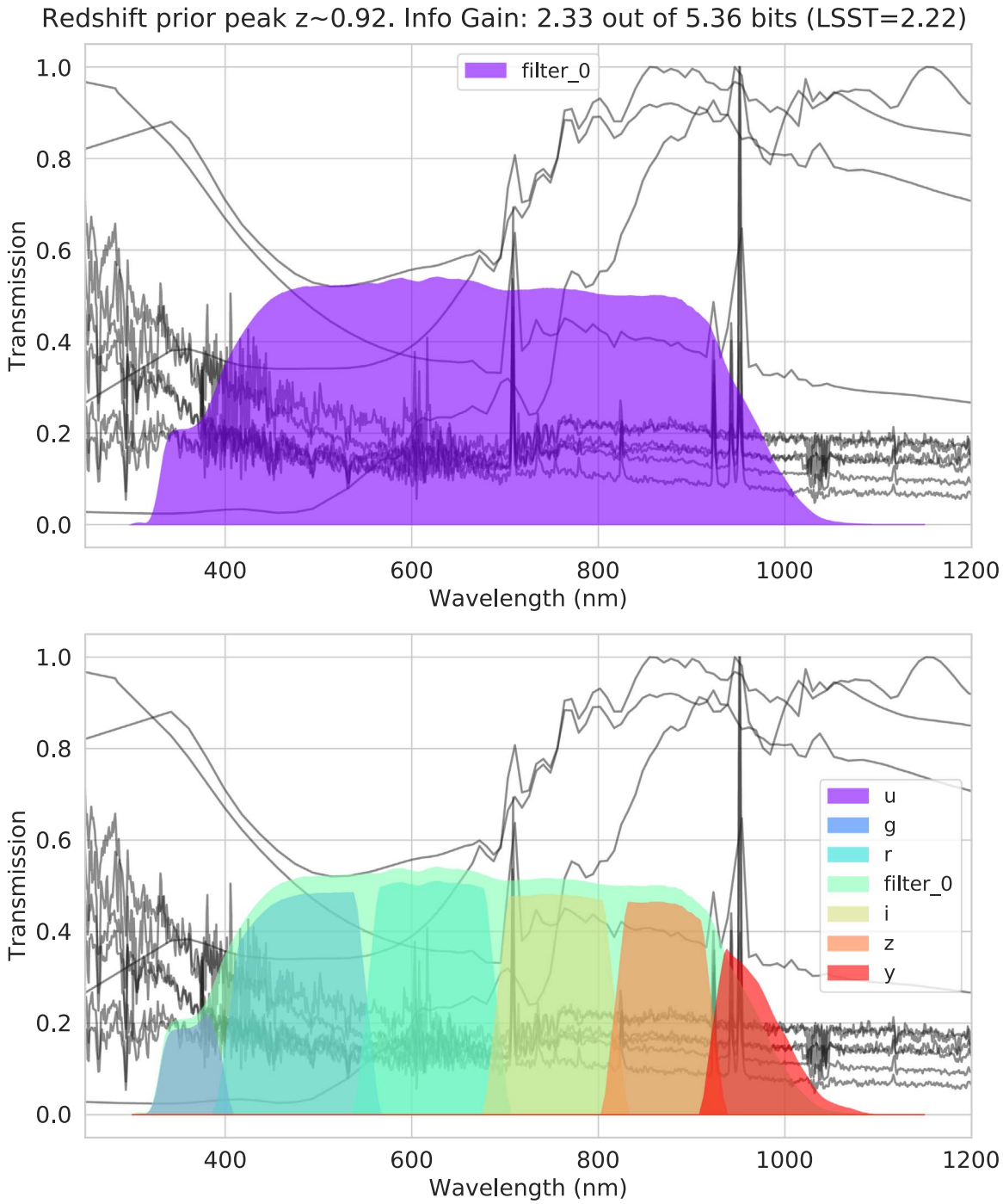

Figure 8. Top: the best additional filter added to LSST filters is a wide filter overlapping all the original LSST filters when the template flux normalized to LSST $i = 25.3$. The CWW–Kinney templates are shown in the background redshifted to the peak of the prior distribution ($z \sim 0.92$). Bottom: additional filter with the LSST filters provided for comparison.

Table 1
Number of Visits to a Field in the LSST Survey for Each Filter

Filter	u	g	r	i	z	y
No. of visits	56	80	184	184	160	160

implementation. In practice, even this resolution can produce artifacts due to insufficient sampling of the distributions.

However, the calculation of this integral can be done by probabilistically sampling from the color distributions. We start from Equation (11) in combination with the approximation $\int p(x)q(x)\,dx = \lim_{N\to\inf} \frac{1}{N}\sum_{x_i\sim p(x)} q(x_i)$ and calculate the following in our code:

$$H(Y|X) = -\frac{1}{N}\sum_i \sum_{x_j\sim p(y_i,x_j)}^{n_i} \log_2\left[\frac{p(y_i, x_j)}{p(x_j)}\right]. \quad (12)$$

To evaluate Equation (12), we draw $N = 10^6$ points from the prior distribution for redshift $p(y_i)$. This gives us n_i points that fall into each redshift bin that we then use to calculate the inner

sum for that bin. For each point in the redshift bin we randomly pick an SED with a uniform probability (a simplification we discuss modifying in future work in Section 9). We then draw a vector of colors from the multivariate Gaussian distribution that models the observed photometric color and uncertainties for the redshifted SED. We save all these color points so that we have a representation of the complete color space for that redshift based on the available galaxy SEDs.

To calculate the sum over the logarithm, we need to find the values for $p(y_i, x_j)$ and $p(x_j)$, where x_j is a point in color space. The value for $p(x_j)$ will be the sum of values measured at x_j from the multivariate Gaussians that are the probability density functions for the colors for each SED at each redshift. To get $p(y_i, x_j)$, this calculation includes only the redshifted SEDs at the specific redshift y_i. We use this technique to sum over the points in each redshift bin and then sum over the values for each redshift bin before normalizing by $\frac{1}{N}$ to get a final answer for $H(Y|X)$. This value for conditional entropy is then subtracted from the full entropy to get the IG.

7. A Realistic Sample

The real world is not nearly as clean as the simple situations discussed above. Rather than observing galaxies of a single spectral type, we observe many different galaxies with different intrinsic spectral characteristics at many different redshifts. Rather than having a single color associated with each redshift, we have a broad distribution of colors associated with each redshift.

To study this, we need a representative set of spectra that evenly samples the expected space of observations. Since we are interested in the estimation of photometric redshifts, we use template sets from Coleman et al. (1980) (CWW) and Calzetti et al. (1994) (Kinney–Calzetti Atlas) supplemented by Arnouts et al. (1999) at UV and IR wavelengths with the GISSEL code (Bruzual & Charlot 1993, 2003). The colors and photometric uncertainties for the Gaussian distributions in color space are calculated based on normalizing all the SEDs to $i = 25.3$ and using a sky background set at $i = 20.47$ with an LSST-like telescope over a 10 yr LSST-like survey. The sky background is modeled with a sky SED provided in the LSST Sims throughputs package (Connolly et al. 2014). We choose $i = 25.3$ for our normalization since this matches the faint limit on the definition of LSST "gold sample" galaxies for photometric redshifts (LSST Science Collaboration et al. 2009). We also use a prior function on redshift that we derive from the photo-z training catalog we describe in Section 8.1. This prior came from fitting a function of a similar form to Equation (6) to the 39,952 training set galaxies with $24.75 < i_{\rm mag} < 25.0$ and is designed to approximate the galaxies expected near the faint normalization of the SEDs. The histogram of the galaxies and the prior are shown in Figure 7. In our tests, we bin the redshift every 0.05 between $0.0 \leqslant z \leqslant 2.3$, giving us 46 total bins. We set the limit at 2.3 because the CWW–Kinney templates' blue limit starts to pass into the bluest wavelengths of our filters at this redshift.

Finally, instead of sampling along a grid of set widths and centers as we did in the example problems, we use scikit-optimize (Head et al. 2018) to optimize the locations and shapes of our filters. Scikit-optimize is an open-source python-based Bayesian optimization package designed to optimize complex spaces such as the high-dimensional IG space in our problem. We use the Gaussian process–based estimator provided in the code to model the output space and choose locations for optimization. In each run, we initialize the space with 10 points before running the optimization and allow the optimization to run in parallel, updating after running a set number of points independently each time.

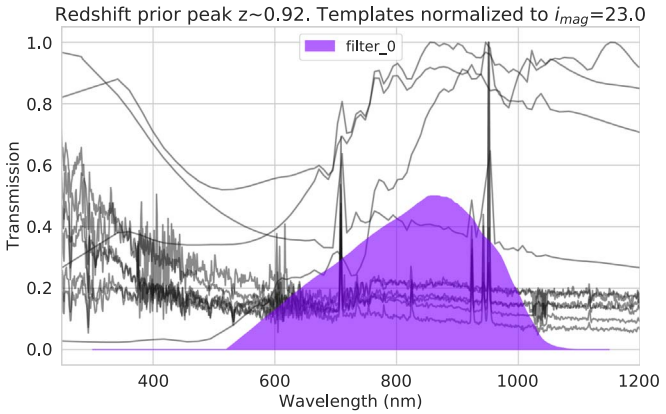

Figure 9. Best additional filter added to LSST filters when the template flux normalized to LSST $i = 23.0$. The filter narrows to focus on the region of the Balmer break.

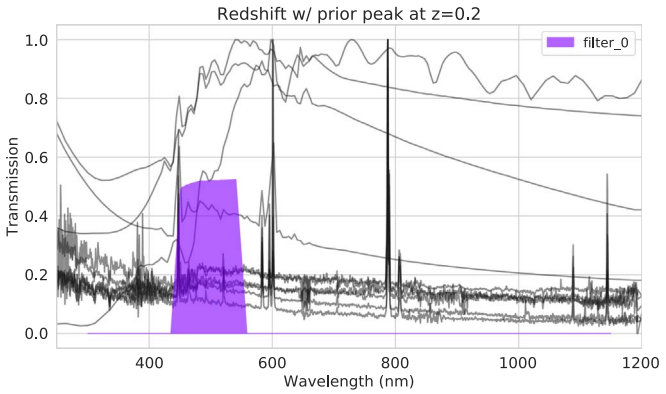

Figure 10. Best additional filter when using a redshift prior distribution that peaks at $z \sim 0.2$. The optimal filter is shifted further toward the blue end of the optical range to get information from the Balmer break around the peak redshift of the redshift prior. The SED templates are shown in the background redshifted to $z = 0.2$.

7.1. Adding a New Filter to LSST

In our first experiment we used the LSST filters as a set of fixed filters and wanted to find the optimal additional filter in the optical range that would benefit photometric redshifts the most. For our simulation we gave this filter an equal number of visits as proposed for the LSST y filter and kept the same number of visits for the other filters, in effect extending the baseline LSST survey shown in Table 1. We allowed the four corners of a trapezoidal filter to move independently in the wavelength range between 300 and 1100 nm. This gave us an optimization with four degrees of freedom that allowed the location and width of the filter to vary, as well as the slope of the wings of the filter on each side.

The resulting filter is shown in Figure 8 with and without the accompanying LSST filters. The best filter is a large filter with wide wings at the blue and red ends. This filter raises the IG

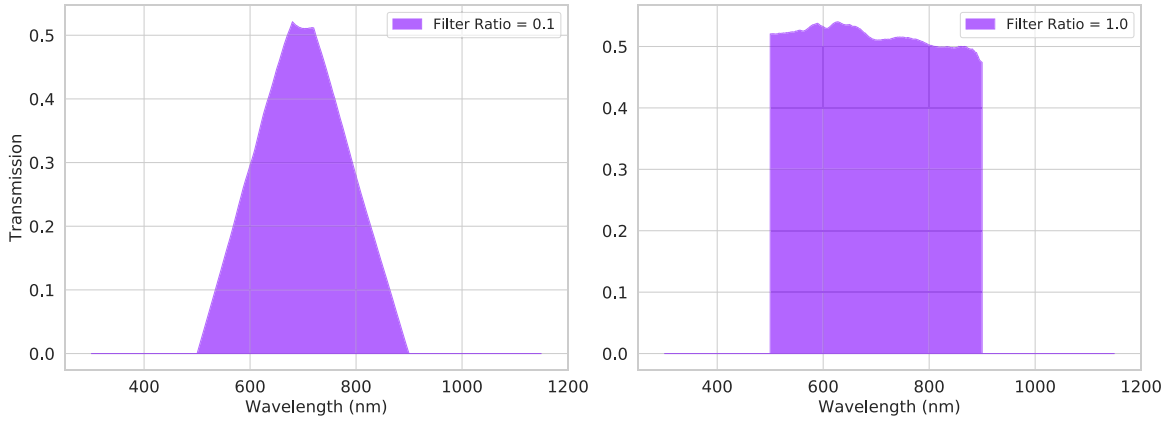

Figure 11. Comparison of the allowable filter shapes. Left: filter with a ratio of top width to bottom width of 0.1. Filters with lower ratios are more triangular. Right: filter with a top-to-bottom width ratio of 1.0. Filters with higher ratios are more rectangular or "top hat" like.

only slightly from 2.22 bits for the LSST filters alone to 2.33 bits.

This wide filter is centered around the Balmer break at the peak redshift of the prior distribution at $z \sim 0.92$. This is obvious in the top panel of Figure 8, where the test SEDs are shown redshifted to this peak value. This seems to confirm what we saw in the examples of Sections 4 and 5. Another thing to notice is that the additional IG provided by a seventh filter to the LSST in the optical range is only a 5% improvement. This indicates that it is difficult to improve the LSST filters by adding wide filters in the optical range.

To understand what is driving the design of the seventh filter, we ran the same experiment but normalized our templates to LSST $i = 23.0$. In this case, shown in Figure 9, the filter narrows to focus on the region around the Balmer break. With brighter galaxies we are able to get a higher-S/N measurement of the wavelengths around the Balmer break at the peak of the prior with a narrower filter. This indicates that the original, wider seventh filter is a result of trying to maximize the S/N of the faint galaxy populations over improving the redshift sensitivity through the introduction of a narrower filter.

7.1.1. Effect of Changing Prior

To verify that the Balmer break is the primary source of information, we reran the optimization with a different prior to see how the location of the seventh filter changed. We created a toy prior that peaks at $z \sim 0.2$. The outcome is shown in Figure 10 and confirms the shift in the filter location toward the location of the Balmer break at the new peak of the redshift prior. Thus, we observe that filters will constrain redshift the best if they can optimally constrain the location of the Balmer break as it moves across the optical wavelength range.

7.2. Six Filter Survey: Properties of Optimal Filter Sets for Photometric Redshifts

The locations and shapes of photometric filters affect the colors observed for stars and galaxies. Photometric redshifts rely on the design of filter systems that will pick up the spectral features for galaxies in the relevant redshift range of a survey. The colors produced by a photometric system are also important for estimating stellar properties (Lenz et al. 1998) and quasar selection (Peters et al. 2015). Here we investigate optimal shapes and locations for photometric redshifts but plan

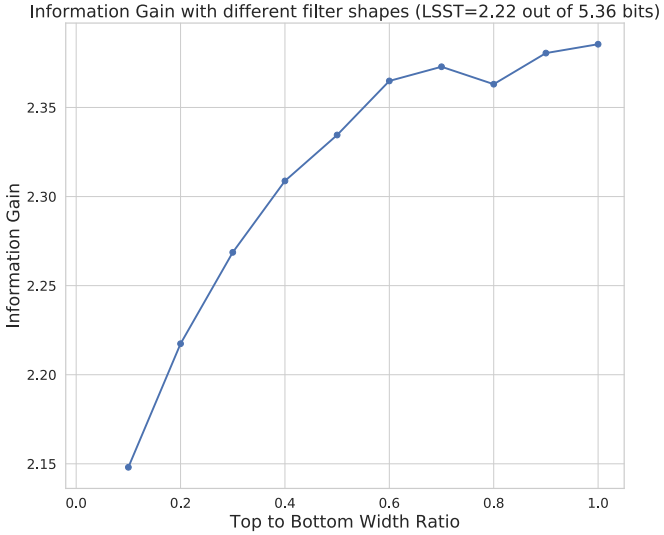

Figure 12. Best IG for a set of trapezoidal filters as a function of the ratio of the width for the top of the filter transmission curve to the bottom width.

to extend this evaluation to other astronomical problems that require colors in future work.

In this test we ran 10 sets of filter optimization, allowing the width and locations of six filters to vary for a total of 12 degrees of freedom. In each test we set a different ratio from 0.1 to 1.0 for the top-to-bottom width of a trapezoidal filter. Figure 11 compares the allowable shapes for the filters, with the most triangular filter having a ratio of 0.1 on the left compared to the most rectangular on the right with a ratio of top width to bottom width of 1.0. We then found the best IG for each filter shape at the end of the optimization run and compared the best values for each width ratio. The results are shown in Figure 12 and discussed in the sections below. Our best IG for six filters was 2.39 bits, which is an increase of 0.17 bits or 7% compared to the 2.22 bits of IG when using the LSST filters. The best performing filters are shown in the top panel of Figure 13, with the LSST filters shown in the bottom panel for comparison.

7.2.1. Filter Shape

The results in Figure 12 clearly show a general trend that increasing the steepness of the wings of a trapezoidal filter leads to better IG up to a ratio of 0.9, where the trend flattens

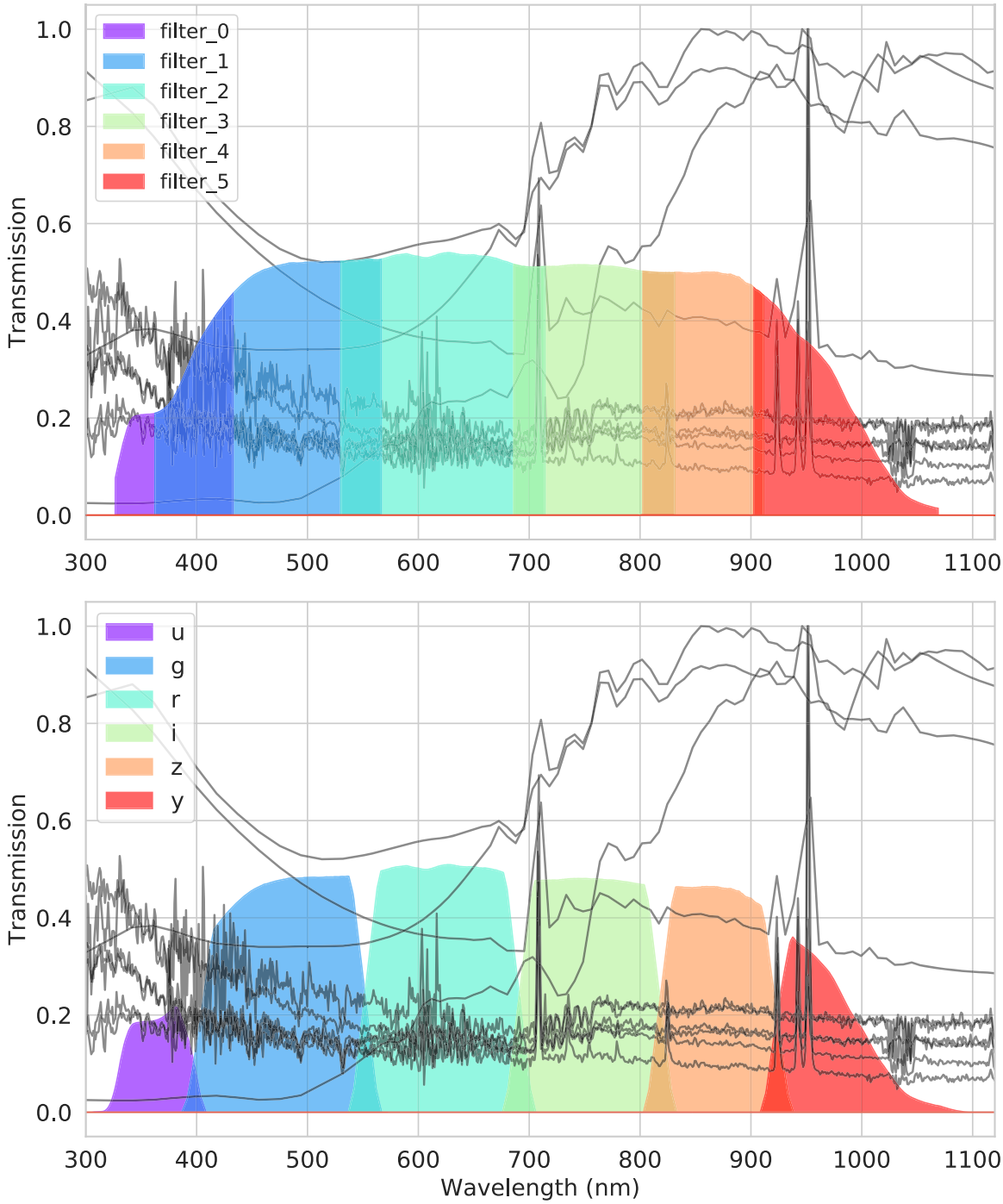

Figure 13. Top: best set of six filters based on our optimization runs. These trapezoidal filters have a top-to-bottom width ratio of 1.0 and an IG of 2.39 bits compared to the 2.22 bits of IG from the LSST filters. Bottom: LSST filters shown for comparison.

out. This overall trend makes sense since the IG is related to the width of the distributions as we saw in Figure 5. There, the better IG came when the possible distribution of colors for a given redshift was narrower. This means that the width of the color distribution is affected by the S/N of the magnitude measurement in each filter. Allowing a wider top of the filter increases the overall transmission for filters of a similar width and thus increases the S/N of the flux measurement in that filter for a given spectrum at a given brightness.

Wider tops to the filters also avoids gaps in between filters without the need for a lot of overlap in the wings of each filter. Gaps in the filters allow strong features to fall between filters and waste information that would otherwise be available. Preventing gaps is necessary to avoid this, but as we will explain in Section 7.2.2, some overlap is beneficial, but there is a limit. Narrow filter wings avoid the information loss caused by filter gaps while minimizing the extra amount of filter overlap that provides redundant information.

7.2.2. Filter Overlap

Filter sets with overlap perform better since overlapping adds information as to the position of a spectral feature in a

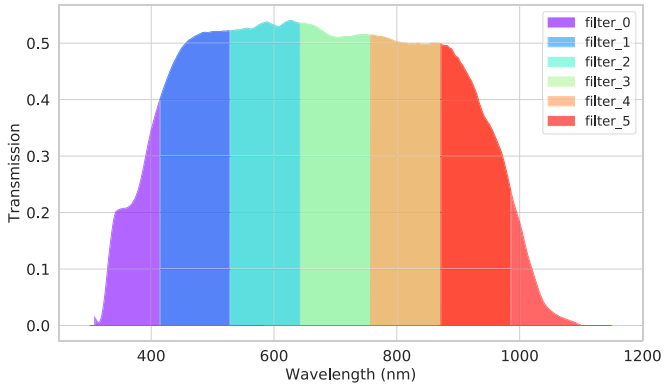

Figure 14. Six filters with 50% overlap of each adjacent filter. The IG for this situation is only 1.99 bits out of 5.36 possible compared to the 2.39 bits gained with the ideal filter set that has the same top-to-bottom ratio of 1.0.

filter. If filters do not have any overlap, then it is harder to distinguish at what redshift a feature in the spectrum passes from one filter to another. In our optimal filter set every filter overlaps with its neighbors.

However, too much overlap creates redundant information and stops being beneficial. We set up an extreme situation with a top-to-bottom ratio of 1.0 just like the optimum filters but with an overlap of half of each filter width and shown in Figure 14. In this setup every wavelength has coverage in two filters. When we calculate the IG for this situation, it has fallen compared to the optimal filter situation above from 2.39 bits to 1.99, demonstrating that complete overlap in every possible wavelength is not ideal.

8. Simulated Photometric Redshift Estimation

To relate the improvement in the IG over the LSST filters to photometric redshift performance, we created a simulated catalog with magnitudes measured for the LSST and new filters.

8.1. Simulated Catalog

We generated simulated catalogs of a circular area on the sky with a radius of $0.^{\circ}8$ using the LSST Catalog Simulations (CatSim) code (Connolly et al. 2014). We generated three different catalogs, one each for the different filter sets: LSST only, six new filters, and LSST+1 filter designs. The LSST CatSim code generates galaxy photometry using templates from Bruzual & Charlot (2003). We ran the code with the different filter sets over the same simulated footprint over a simulated 10 yr survey with the same survey properties as given in Graham et al. (2018). Where we included a seventh filter we gave it 160 visits to match the number in the LSST y filter. Following the same procedure as Graham et al. (2018), we included fainter galaxies and used the simulated magnitude errors to apply a random normal scatter to the catalog before making a cut at LSST $i < 25.3$. Then, we made a final cut and only kept objects with a redshift $\leqslant 2.3$ since this was the range of our redshifts when optimizing the filters in Section 7. This final cut was then split to give us 61,484 test objects and a training catalog with 301,400 objects in our simulated catalogs.

8.2. Calculating Photometric Redshifts

We used the Color Matched Nearest Neighbors (CMNN) redshift estimation code of Graham et al. (2018) on our simulated data. The CMNN photo-z code calculates the Mahalanobis distance in color space between each test galaxy and galaxies in the training catalog. The photo-z value for the test galaxy is then the redshift of the nearest neighbor in the training catalog. The CMNN photo-z code is designed to enable rapid characterization of the relative performance of photo-z estimation for catalogs with different filter designs (rather than the absolute photo-z accuracy). A comparison of the accuracy of CMNN compared to other template-fitting and machine-learning-based techniques is presented in Schmidt et al. (2020). In this paper the accuracy of CMNN, for the metrics we use in this paper, was comparable to or better than standard template-based approaches.

We ran the CMNN photo-z code on each of our three simulated catalogs and compared the results. Figure 15 shows the density plots comparing the input catalog redshifts to the photometric redshifts. Between the two LSST-based filter schemes there does not appear to be much difference in the density plots. The six new filters do seem to improve the results at redshifts $z < 0.6$, where there is a clear increase in photo-z scatter for $0.2 < z < 0.6$ visible in a cross-like feature in the density plots. Beyond this the density plots once again look similar to the LSST, except there does seem to be more scatter at redshifts greater than 1.5, where the Balmer break leaves the optical range. Since we have previously shown that the IG is strongly affected by the Balmer break, this is not surprising.

To get a more informative look at the errors, we use the photometric redshift error defined in Graham et al. (2018). The photometric redshift error is defined as $\Delta z_{1+z} = (z_{\rm true} - z_{\rm phot})/(1 + z_{\rm phot})$, and we use this to calculate four performance metrics. To analyze the error, we plot the bias (mean Δz_{1+z}) and standard deviation of the Δz_{1+z} values as a function of the true redshift. We plot a robust standard deviation, which is the standard deviation of the interquartile range of the errors multiplied by 1.349 to create a value comparable to a standard deviation. Finally, we plot the fraction of outliers with Δz_{1+z} values greater than 0.15. We use 12 bins in the redshift range from 0 to 2.3 and plot the values in Figure 16. In Figure 17 we compare the differences to the LSST values for each new filter set. The dashed black line is set where the values are equal to the LSST, so that above this line the new filter set is worse and below the line the new filters perform better than LSST.

The six new filters obtained through IG optimization do offer more improvement compared to adding a seventh filter, which is consistent with the greater IG improvement. Overall the six new filters improve the standard deviation by 7.1% and the outlier fraction by 9.9%. These gains are driven by improvements for the redshifts below 0.9 and traded for performance losses at $z > 1.5$. This was noted in the density plots and is consistent with IG focusing on the presence of the Balmer break in the optical range.

Adding a seventh optical filter to the LSST filters offers only slight benefits. There is a slightly lower overall standard deviation around the $z \sim 0.9$ peak of the prior distribution we used, and over the whole test set we have a 1.6% improvement in standard deviation. The biggest gains are a 4.6% improvement in overall outlier fraction and a 16.6% improvement in the bias, with gains appearing to be spread out through the redshift range, unlike the six new filters.

We also performed the photo-z analysis with our alternate seventh filter from Section 7.1 that was optimized with brighter templates. When comparing the photometric redshift performance between the narrower and wider filters, we found that

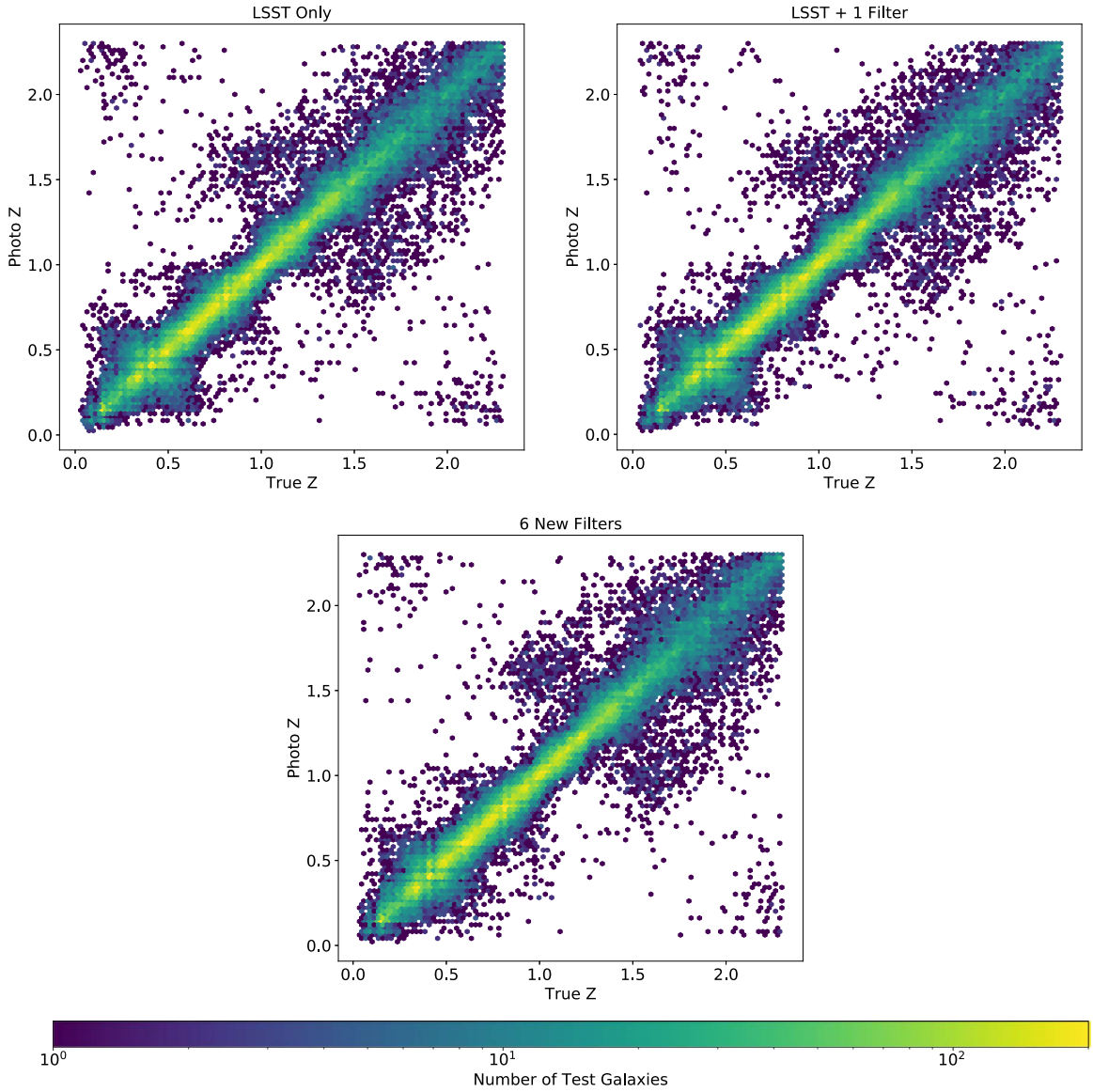

Figure 15. Density plots for the results from photometric redshift estimation on the simulated catalogs with the CMNN photo-z code and the different filter sets. Top left: LSST filters only. Top right: LSST + one new filter. Bottom: six new optimized filters.

for galaxies $i < 23$ the fraction of photo-z outliers is 4.7% worse for the wider filter and for $i > 23$ the outlier fraction is 2.3% better for the wider filter. For the full population of galaxies the wider filter has a 2.1% smaller fraction of photo-z outliers. These results are consistent with our claim in Section 7.1 that the width of the seventh filter is driven by the faintness of the galaxies we targeted with our optimization.

9. Discussion

We were only able to provide small improvements over the LSST filters in terms of both IG and actual photometric redshift estimation. This is because the LSST filters already have similar features to those we identified as optimal for photometric redshift filters. The LSST filters have no gaps between them, and each filter has a small degree of overlap with the adjacent filters. The filters are also nearly top hat in shape with slight wings on each side, which is consistent with our findings of optimal filter sets. Our results from adding a seventh filter in the current LSST wavelength range show that another filter in optical wavelengths is not a good way to improve photometric redshifts with LSST. In fact, since we have shown that the Balmer break is so important, it seems that a greater improvement to LSST photometric redshifts would come from following the break into the infrared wavelengths. This is in line with previous work highlighting the potential photometric redshift improvement from combining LSST observations with infrared data from future space telescope missions (Jain et al. 2015; Rhodes et al. 2017; Graham et al. 2020). In a future paper we will look at how much information is gained from adding infrared and UV filters, as well as what optimal filters in these wavelengths designed to complement LSST would look like.

The practical application of the optimal filters in our work shows that our method has merit and can be used to design observations tailored to photometric redshift estimation. Increasing IG did correlate to an improvement in photo-z performance. However, limiting ourselves to redshifts up to $z = 2.3$, we are not able to provide insight into how filters can improve the Lyman versus Balmer break degeneracy. This is a

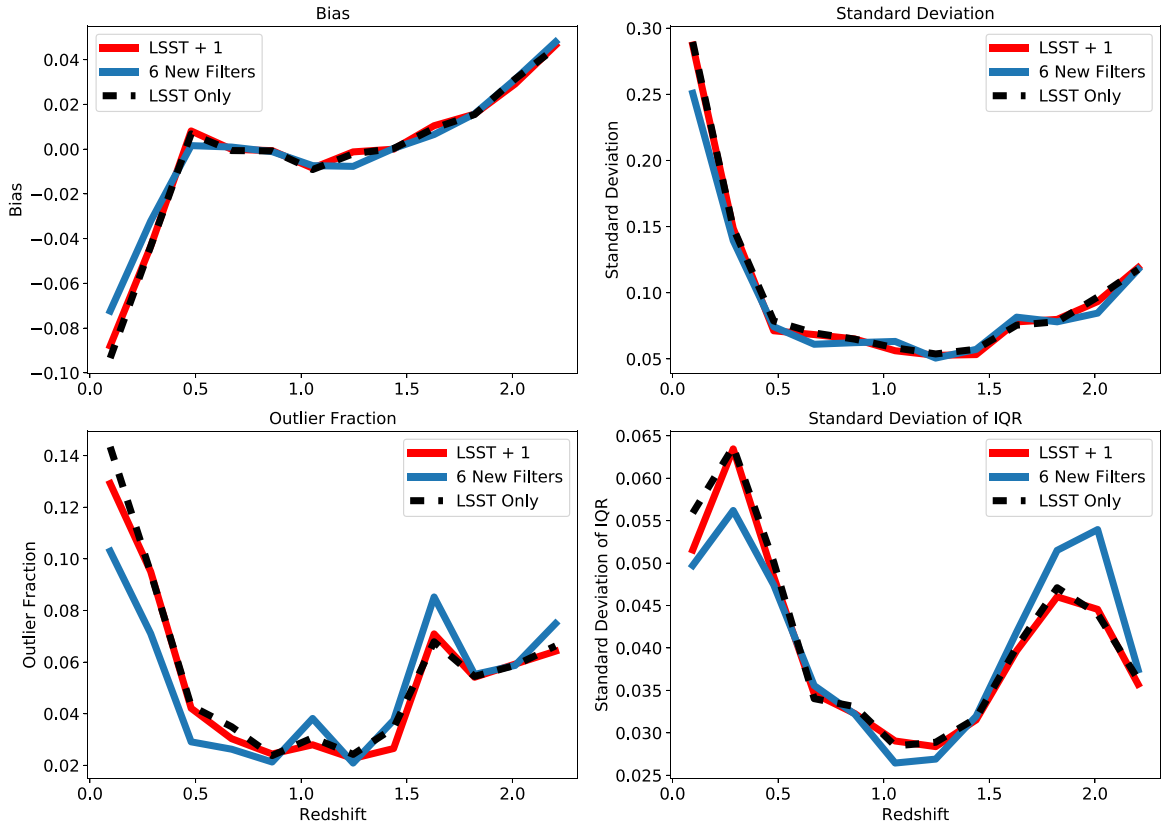

Figure 16. Comparing the photometric redshift errors of the three different filter sets. As expected from the density plots, adding a new filter to LSST does not change much, and the six new filters reduce outliers at low redshift but trade this for performance at higher redshifts.

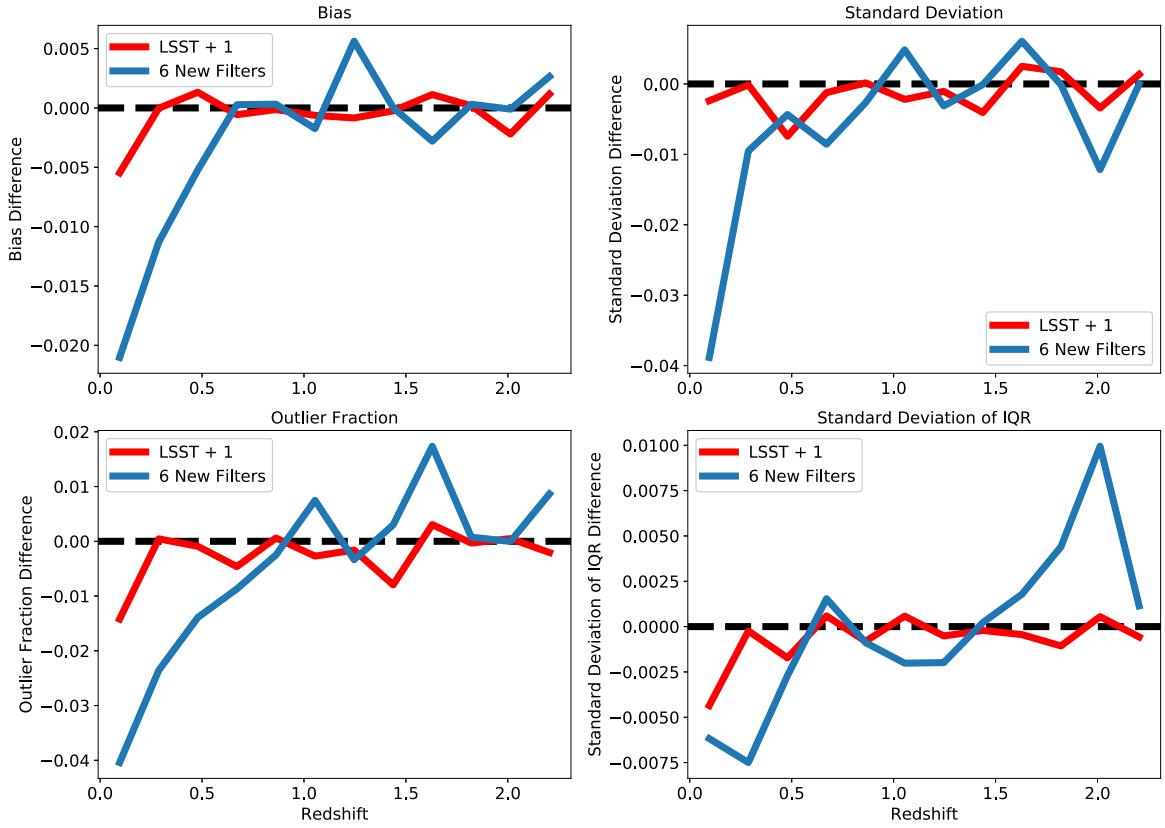

Figure 17. Comparing the differences in photometric redshift errors of the two new filter sets to performance with the LSST filters only. The black line indicates that errors are the same as the LSST filters. Below the black line means improvement over LSST, while above the black line indicates worse performance. The added filter seems to slightly improve bias and overall standard deviation around the peak of the redshift prior at 0.9 and helps reduce outliers overall. As noted above, the six new filters outperform LSST at lower redshifts in return for slightly degraded performance at higher redshifts.

large problem in photometric redshift estimation, and exploring filter design with our method at a larger redshift range could provide interesting insights into this. Gathering templates that apply to the blue end of the optical range beyond $z = 2.3$ and to higher redshifts will help us explore the question in the future.

We used a simple redshift prior in this work, but many options to enhance the priors we use exist. For example, instead of sampling each template with a uniform probability, we could include a prior to weight certain templates of galaxy types more heavily at different redshifts. More descriptive priors reduce the starting value of entropy for filter optimization and thus change the amount of IG possible from the filter design. In our catalog used in Sections 7 and 8 we had 46 redshift bins. With a flat prior where every redshift is equally likely, this equals 5.52 bits of entropy, representing the maximum level of uncertainty in the problem. With the prior based on the training data we reduced this to 5.36 bits of entropy. One way of looking at this is that the prior itself provides 0.16 bits of IG. Then, our best set of filters was able to reduce the entropy a further 2.39 bits. Introducing more advanced priors could help tailor our code to produce filters truly optimized for the practical application of photometric redshift estimation. In addition, optimal photo-z filters might be different from the broadband filters we used in this work. We could look at how a large number of narrowband filters or a comb filter would be optimized for photometric redshifts. Or we could move away from trapezoidal filters and allow more complex shapes in the design of an individual filter.

We applied our methodology to galaxies and photometric redshifts, but we can easily apply it to any set of templates to find the ideal filters that will differentiate between the corresponding astronomical objects. Future work will focus on applying the IG methodology to design filters that optimize properties beyond photometric redshifts such as stellar observations or quasar selection. For example, templates for different stellar types could be used to design filters that optimize observations to determine stellar properties.

10. Conclusion

We have introduced a new technique to apply information theory to the design of filters in order to optimize photometric redshifts. We showed its theory and provided insight into its use with three simple examples before using it in a practical situation. We created an optimal set of six filters to cover the optical wavelengths in an ideal manner for photometric redshifts. This application revealed the general attributes of an ideal filter set for photometric redshifts. Ideal filters will have narrow wings and be near top hat in shape. They will also have a small amount of overlap. We showed that the main information for photometric redshift estimation comes from the Balmer break, and optimal filters will focus on maximizing IG of this break at the peak of the redshift distribution.

We applied the two different filter sets to a simulated catalog of a sample of photometric data and compared the photometric redshift estimation results to the LSST filters. We showed that a set of six filters optimized using IG could improve the standard deviation of the errors associated with photometric redshifts by 7.1% overall at redshifts up to 2.3 over the LSST filters and outliers up to 9.9%, but improved performance at lower redshifts was traded for slightly worse results than LSST at higher redshifts. The LSST filters perform near the optimal set and have features we identify as optimal for photo-z filters, such as overlap with neighboring filters and a nearly rectangular shape. We also discuss future directions that will improve our technique and will be possible with improvements to the code we used in this work. This python code, SIGgi, is publicly available at https://github.com/dirac-institute/siggi and is pip installable.

This work was supported by the U.S. Department of Energy, Office of Science, under award No. DE-SC-0011635. J.B.K. and A.J.C. also acknowledge support from the DIRAC Institute in the Department of Astronomy at the University of Washington. The DIRAC Institute is supported through generous gifts from the Charles and Lisa Simonyi Fund for Arts and Sciences and the Washington Research Foundation.

Software: scikit-optimize (Head et al. 2018), SIGgi v0.1.6 (Kalmbach 2019), Scikit-Learn (Pedregosa et al. 2011), LSST Simulations Software Stack (Connolly et al. 2014), Numpy (Oliphant 2006), Matplotlib (Hunter 2007), Scipy (Virtanen et al. 2019), Pandas (McKinney 2010).

ORCID iDs

J. Bryce Kalmbach https://orcid.org/0000-0002-6825-5283
Jacob T. VanderPlas https://orcid.org/0000-0002-9623-3401
Andrew J. Connolly https://orcid.org/0000-0001-5576-8189

References

Adami, C. 2004, PhLRv, 1, 3
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Baum, W. A. 1962, in IAU Symp. 15, Problems of Extra-galactic Research, ed. G. C. McVittie (New York: Macmillan), 390
Bruzual, A., & Charlot, S. 1993, ApJ, 405, 538
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Calzetti, D., Kinney, A. L., & Storchi-Bergmann, T. 1994, ApJ, 429, 582
Cincotta, P. M., Mendez, M., & Nunez, J. A. 1995, ApJ, 449, 231
Coleman, G. D., Wu, C.-C., & Weedman, D. W. 1980, ApJS, 43, 393
Connolly, A. J., Angeli, G. Z., Chandrasekharan, S., et al. 2014, Proc. SPIE, 9150, 915014
Connolly, A. J., Csabai, I., Szalay, A. S., et al. 1995, AJ, 110, 2655
Graham, M., Connolly, A., Wang, W., et al. 2020, ApJ, Submitted
Graham, M. J., Drake, A. J., Djorgovski, S. G., et al. 2013a, MNRAS, 434, 3423
Graham, M. J., Drake, A. J., Djorgovski, S. G., Mahabal, A. A., & Donalek, C. 2013b, MNRAS, 434, 2629
Graham, M. L., Connolly, A. J., Ivezić, Ž, et al. 2018, AJ, 155, 1
Head, T., MechCoder, Louppe, G., et al. 2018, scikit-optimize v0.5.2, Zenodo, doi:10.5281/zenodo.1207017
Huijse, P., Estévez, P. A., Förster, F., et al. 2018, ApJS, 236, 12
Hunter, J. D. 2007, CSE, 9, 90
Ivezić, Ž., Tyson, J. A., Acosta, E., et al. 2019, ApJ, 873, 111
Jain, B., Spergel, D., Bean, R., et al. 2015, arXiv:1501.07897
Kalmbach, J. B. 2019, dirac-institute/siggi: Siggi v0.1.6, Zenodo, doi:10.5281/zenodo.3267967
Koo, D. C. 1985, AJ, 90, 418
Kullback, S., & Leibler, R. A. 1951, Ann. Math. Stat., 22, 79
Lenz, D. D., Newberg, J., Rosner, R., Richards, G. T., & Stoughton, C. 1998, ApJS, 119, 121
Lindley, D. V. 1956, Ann. Math. Stat., 27, 986
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
McKinney, W. 2010, in Proc. 9th Python in Science Conference, ed. S. van der Walt & J. Millman, 51, http://conference.scipy.org/proceedings/scipy2010/mckinney.html
Oliphant, T. 2006, NumPy: A Guide to NumPy (USA: Trelgol Publishing)
Ormos, M., & Zibriczky, D. 2014, PLoSO, 9, e115742
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Peters, C. M., Richards, G. T., Myers, A. D., et al. 2015, ApJ, 811, 95

Pogosian, L., Corasaniti, P. S., Stephan-Otto, C., Crittenden, R., & Nichol, R. 2005, PhRvD, 72, 103519
Rhodes, J., Nichol, R. C., Aubourg, É, et al. 2017, ApJS, 233, 21
Seehars, S., Amara, A., Refregier, A., Paranjape, A., & Akeret, J. 2014, PhRvD, 90, 023533
Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, MNRAS, submitted (arXiv:2001.03621)
Shannon, C. E. 1948, BSTJ, 27, 379
Virtanen, P., Gommers, R., Oliphant, T., et al. 2019, arXiv:1907.10121
Weir, N., Fayyad, U. M., & Djorgovski, S. 1995, AJ, 109, 2401