Analysis of Emmision Lines of Galaxy NGC6240

Amy Hoffman

February 28, 2019

Abstract

The NCG6240 galaxy is filled with light of wavelengths. The shape of the spectral data of normal light appears to look continuous. However, the release of hot gas creates emission lines, or spikes, in the seemingly continuous data. The identification of the cause of the emission line depends on determining the intensity of the emission line. Using a linear regression line, the data points of the emissions lines were separated from the normal light in the galaxy. Then, using first order approximations of the first derivative, the emission line peaks and the nearest local minimums on either side of the emission peak were identified. The intensity of the emission line is measured by the area under the emission line that lies above the normal light intensity. The normal light intensities were approximated by the area of a trapezoid or rectangle depending upon whether the emission lines overlap. This method suggested the intensities of the emission lines with peaks at wavelengths 3882, 6444, 6724, 6746, and 6886 Angstrom (A) have intensities of 29.2628, 25.1084, 48.7688, 47.3373, 24.0289 W/Am^2, respectively. The intensities are accurate to 1e-3, as the trapezoid rule and the midpoint Riemann sum numerical integration produce approximate intensities varying less than 1e-3.

Contents

1 Introduction **2**

2 Identification of Outlier Data **2**

2.1 Identification of Peaks 2

3 Identification of Integration Bounds **4**

4 Numerical Integration **4**

5 Conclusion **7**

---

1 Introduction

The provided data set contained 1726 spectral data entries of the wave intensity (measured in Watts per square meter per Angstrom W/Am^2) at various wavelengths (measure in Angstroms A) in the galaxy. The light from stars in the galaxy typically creates a rather continuous curve, but the existence of hot gas creates emission lines, or spikes in the data. However, the emission lines do not occur at predictable intervals. The goal of the project was to create a systematic process by which the astronomer could isolate the emission lines and determine their intensity. In determining the intensity, the emission line can be paired to a specific substance, as all gases have different intensities. This systematic process involved linear regression lines, numerical first derivatives, and numerical integration.

2 Identification of Outlier Data

Since the light emitted by the rest of the galaxy is smooth compared to the emission lines, the peaks of the five emission lines are outliers, as seen in Figure 1.

Without the emission lines the light intensities appear to be somewhat linear. Therefore, in order to determine the emission peaks a linear regression line was fitted to the data and the outliers in the data would have the greatest error. However, it is important to note that the five points with the greatest errors are not necessarily the five emission line peaks. The error between the actual and estimated intensities of the linear regression line was 0.14 and the regression had mean squared error of 57.46. The fitted linear regression line follows the equation

$$y = 0.0002x - 0.0192$$

where x is the wave length and y is the intensity.

While the regression line may not fit the data perfectly, as seen in Figure 1, the regression line still provides a good sense of the extreme outliers. Compared to the normal intensities, the emission peaks will still have a greater error, as they are extreme outliers both in the data and in relation to the linear regression predicted values. In order to subset the spectral data into only extreme outliers, the extreme outliers were classified to have an error greater than two standard deviations from the mean, or approximately 0.516. Figure 2 shows the spectral data, regression line, and the data points with an error greater than the threshold.

Too large of an outlier threshold will result in false negatives, and exclude some emission lines. But too small of a number will result in a false positives. If the threshold drops below 0.37, the method identifies false positives. If the threshold is above 0.76, the fails to identify all emission lines. Though threshold values between 0.37 and 0.76 will return the proper emission lines, the proper identification of emission lines is sensitive to the threshold value.

2.1 Identification of Peaks

The linear regression model did a good job of sub-setting the data to only contain data points on the emission lines. While it might be tempting to jump to calculating the emission line intensities, this would be unwise. Notice that two of the five emission lines nearly overlap near 6750 in Figure 1. Therefore, this requires further investigation to determine exactly where one emission line ends and the other begins. Determining the emission line intensities with the current information would likely result in classifying the overlapping emission lines as one emission line.

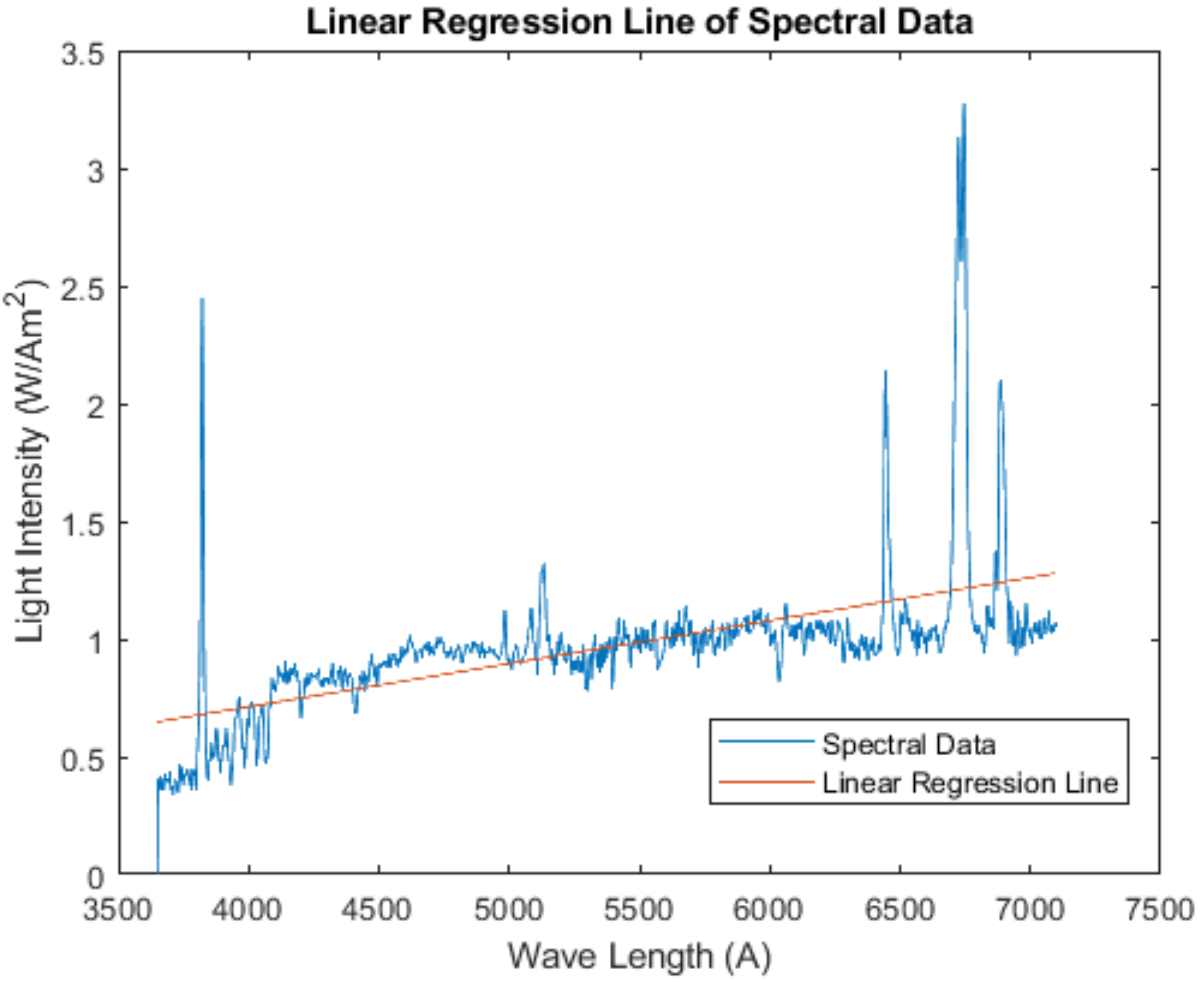

Figure 1: Graph of the spectral data and a fitted linear regression line.

As a result, an extra step was required. The data was then further subsetted into groups of consecutive data points then validated to contain one or multiple emission peaks. This method leverages the fact that the first derivative will change from positive to negative at the peaks. Isolating the peaks was completed in these four steps:

- Find first order first numerical derivative
- Create a boolean vector according to function

$$\begin{cases} 0 & d \geq 0 \\ 1 & d \leq 0 \end{cases} \tag{1}$$

  where d is the first derivative

- Multiply vector by the sign (1 or -1) of the numerical derivative of the next data point in the sequence

This results in a vector consisting of the values -1,0, and 1, where -1 signals a local minimum between peaks, 1 is an emission peak, and 0 is neither. The one downside of this method is that it assumes the peak will not occur at the last data point. Figure 3 shows the identified emission peaks, which occur at 3822, 6444, 6724, 6746, and 6886.

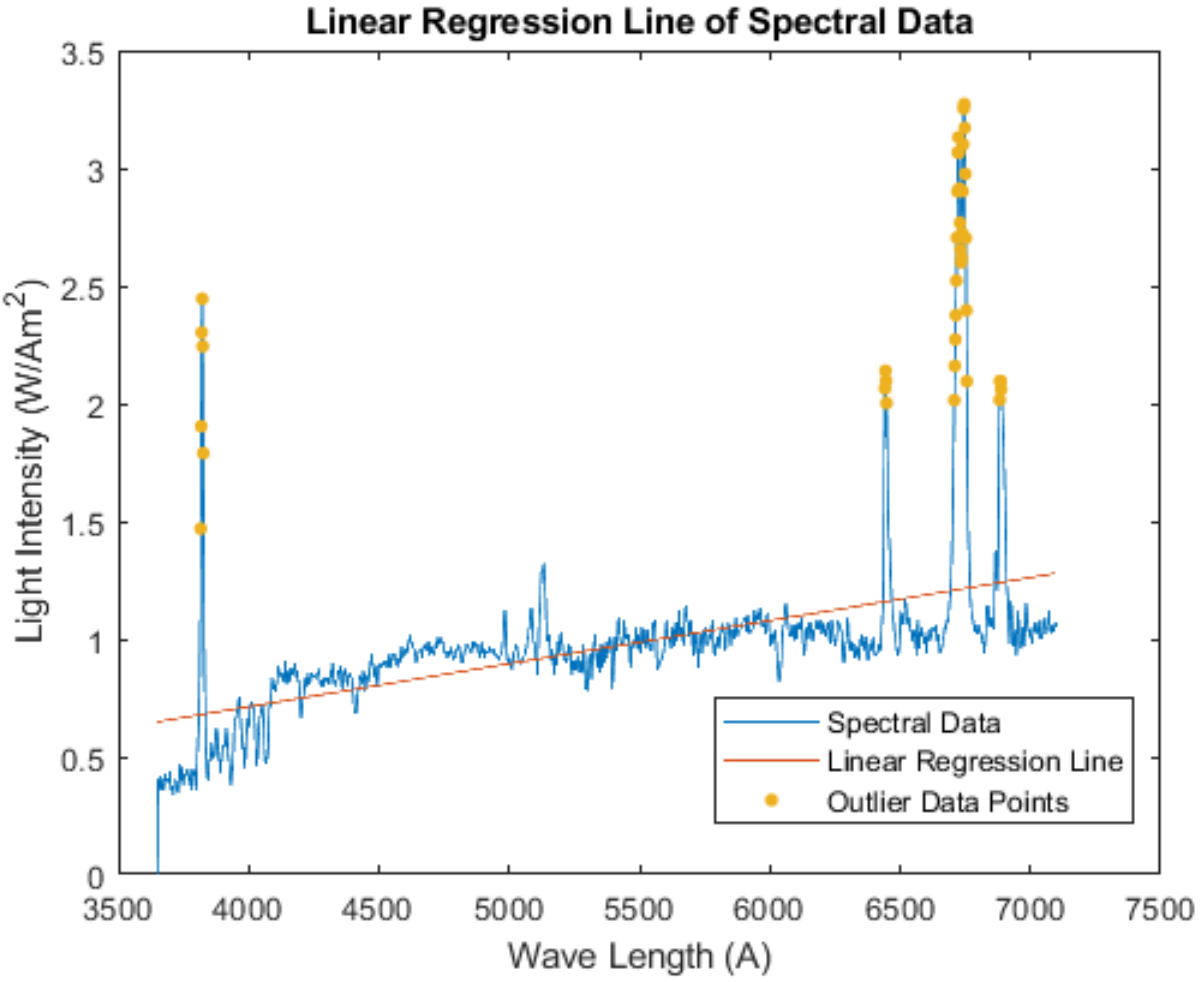

Figure 2: The data points identified as extreme outliers, as their error from the linear regression predicted values were greater than two standard deviations from the mean of the errors.

3 Identification of Integration Bounds

The peaks of the emission lines are now known, but determining the intensity of the emission line requires integrating over the wavelength. Thus, the next step is to determine the wavelength in order to provide bounds for the numerical integration. The bounds of integration for the emission lines are x values of the local minimum closest to the peak on either side. Like isolating the peaks, identifying the bounds of integration depended on the shift from a negative first derivative to a positive first derivative. Using the same method as above, the nearest local minimum to the left and to the right of the emission peak were determined as (x_1, y_1) and (x_2, y_2), where x_1 and x_2 will serve as integration bounds.

4 Numerical Integration

The intensity of the emission line is the numerical integral of the wavelength above the normal light intensities in the galaxy, and thus can be calculated by the area bounded by the emission line, x-axis and the integration bounds minus the intensity of the normal light in the galaxy, as seen in the equation

$$I = \int_{x_1}^{x_2} f dx - \int_{x_1}^{x_2} g dx$$

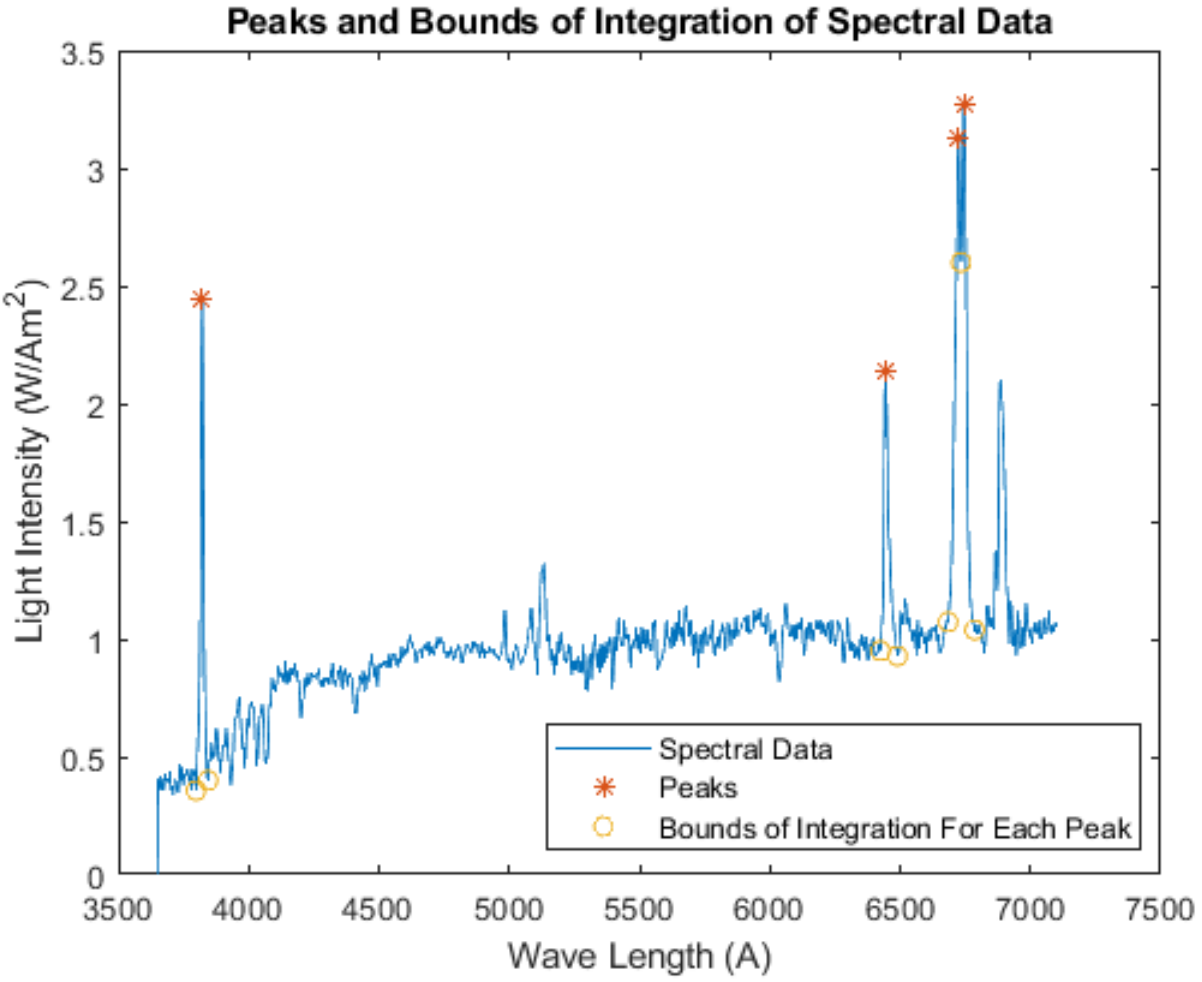

Figure 3: The peaks of the emission lines and the nearest local minimum on both sides. The x values of the local minimums will serve as the integration bounds.

where I is the total intensity, f is the spectral data, and g is the linear threshold that separates the emission intensity from the intensity of the normal light. The threshold g depends upon the proximity of the integration bounds, (x_1, y_1) and (x_2, y_2), and was decided according to the following two cases:

1. If the absolute value of the difference between y_1 and y_2 is greater than the standard deviation of the spectral data intensities, then

$$g(x) = min(\{y_1, y_2\})x.$$

   This function approximates the intensity of the normal light as a rectangle $(x_2 - x_1)$ wide and $min(\{y_1, y_2\})$ tall.

2. If the absolute value of the difference between y_1 and y_2 is less than the standard deviation of the spectral data intensities, then $g(x)$ is the linear line connecting (x_1, y_1) and (x_2, y_2), as modeled by the equation,

$$g(x) = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) + y_1.$$

   This method approximates the normal light intensity as the area of the trapezoid bounded by the linear line, x-axis, and the integration bounds.

The use of two functions for $g(x)$ is important, as over or under estimating the normal light intensity in the galaxy affects the accuracy of the intensity of the emission lines. Table 1 shows the approximated intensities when the normal light of the galaxy is approximated by the area of the rectangle created by Δx and y_1, the area of the rectangle created by Δx and y_2, and the area of a trapezoid formed by the linear connection of (x_1, y_1) and (x_2, y_2). For comparison purposes, Table 1 only shows the intensity approximation using the trapezoid rule.

	Intensity W/Am^2		
Emission Peak Wave Length A	Rectangle y_1 Tall	Rectangle y_2 Tall	Trapezoid
3822	30.2587	28.2669	29.2628
6444	24.3659	25.8509	25.1084
6724	48.7688	-27.6112	10.5788
6746	-33.9023	47.3373	6.7175
6886	18.6889	29.3689	24.0289

Table 1: The trapezoid rule approximated intensities of the emission lines whose peak occur at each wavelength. These approximations show the error caused when always approximating the intensity of the normal light as a rectangle or trapezoid.

Notice the approximate intensities when using trapezoids and when using rectangles to approximate the normal light intensities are fairly similar for all emission lines except the emission lines with peaks at 6724 and 6746. This is because these two emission lines are overlapping. As a result, using a trapezoid or rectangle the height of the integration bound between the emission peaks will result in a gross under estimation, and in some cases a negative value, due to the gross over estimation of the normal light intensity. As a result, the best method for predicting the emission line intensities is a mixed use of trapezoids and rectangles, though the mixed approach does not ensure the exact approximation of the normal light intensities.

However, it is important to note the method of approximating the normal light is not the only contributor to an approximation error. Different numerical integration techniques are more accurate than others, and in some cases more suited to the context of the problem, which can cause a margin of error in approximating the emission line intensities. Table 2 shows the approximate intensity of each emission line using five numerical integration techniques: the trapezoid rule, left Riemann sum, right Riemann sum, midpoint Riemann sum, and Simpson's rule.

	Intensity W/Am^2				
Emission Peak Wave Length A	Trapezoid Rule	Left Riemann Sum	Right Riemann Sum	Midpoint Riemann Sum	Simpson's Rule
3822	29.2628	29.2195	29.3061	29.2628	29.5011
6444	25.1084	25.1309	25.0859	25.1084	25.7198
6724	48.7688	47.2412	50.2964	48.7688	50.4281
6746	47.3373	48.8996	45.7750	47.3373	47.3319
6886	24.0289	24.2514	23.8064	24.0289	24.0313

Table 2: The approximate intensities of the emission lines whose peak occur at each wavelength found using five common numerical integration techniques. These approximations use the mixed approach of trapezoids and rectangles discussed above to approximate the intensity of the normal light in the galaxy.

Notice how the trapezoid rule and the midpoint Riemann sum produced approximations varying by less than 1e-3%. As expected, the left and right Riemann sums yielded approximations with much greater variation, as the few data points and steep slopes of the emission lines lead to large over and under approximations. Further, to no surprise, Simpson's method produced approximations generally greater than those of the trapezoid rule and midpoint Riemann sum. This is because Simpson's rule uses concave downward parabolas rather than linear lines to approximate the emission line intensity. The assumption that the emission lines follow a more linear pattern makes the use of Simpson's rule an inappropriate integration technique for this situation, and the common over and under estimations of the left and right Riemann sums suggests the definite integral has a greater error. Thus, the trapezoid rule and midpoint Riemann sum provide the best estimates of intensities of the emission lines, meaning the margin of error in approximating the intensity as a result of integration technique, when using these techniques, is less than 1e-3%. Yet, it is important to note the midpoint Riemann sum requires more computation in order to calculate the midpoint of a linear line connecting two data points.

5 Conclusion

The development of this method allows the astronomer the identify and classify the cause of emission lines within the noise of the light of the galaxy with ease. The mixed use of trapezoids and rectangles to approximate the intensity of the normal light provides the astronomer with more accurate emission line intensity approximations, and thus allows for a more accurate classification of the the emission line cause. Moreover, this method is scale-able, as the creation of a linear regression line and approximating the first order first derivative is computationally simple. Further, since the trapezoid rule and the midpoint Riemann sum result in very similar approximate intensities, the astronomer could use to the trapezoid rule to further improve computation speeds.