A Deep, Archival Search for Tidal Disruption Events and Rate Constraints

SURF Interim Report 2

Thuwaragesh Jayachandran

Caltech

I. Introduction

Tidal Disruption Events (TDEs) are astronomical phenomena in which a star gets shredded and torn due the tidal forces exerted by a supermassive black hole (SMBH) due to the star getting closer to it than the tidal radius. The resulting multiwavelength radiation causes flares lasting months to years on the time-domain[1]. These flares help detect, observe and study quiescent black holes in distant galaxies.

However, there are several unanswered questions regarding TDEs. The rates of TDEs are yet to be constrained. Also correlations with host galaxy properties, and the factors driving TDEs are also yet to be studied. On top of these, the difficulty in observing TDEs, especially the low luminous ones, suggests that several TDEs might not have been detected in the past[2, 3].

An observed flare could have its origin in an Active Galactic Nucleus (AGN) or a quiescent blackhole. In a quiescent blackhole, the cause is most likely a TDE, however, in an AGN there are several other explanations: supernovae, microlensing, blackhole mergers. Also, AGN light curves tend to exhibit intrinsic baseline variability, as compared to the almost unchanging baselines of a quiescent blackhole. This suggests that TDEs occurring in AGN are less likely to be detected and that there is a possibility that the AGN flares detected in the past could be attributed to TDEs. Therefore it is necessary to study flares of all types in general to get a full understanding of TDEs and their flares.

II. Objective

In this project we aim to address these problems by using re-processed archival data from the Zwicky Transient Facility (ZTF). Utilizing methods used in the past along with our novel approach towards the problem, we are currently working on developing software that can detect and classify flares purely from light curves. At its completion, we aim to use the software to build a catalog of flares, regardless of whether they are due to a TDE or not, from the ZTF data. We will constrain parameters, obtained from distributions of the catalog and those from previous studies, so as to segregate TDE flares from the catalog to obtain a large collection of TDEs, including the faintest detected ones, and study them in depth to further understand the physics surrounding TDEs.

III. Project Outline and Approach

A. Gaussian Processes

A Gaussian Process (GP) is defined as a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution.[7] With the specification of a covariance matrix as the 'kernel' of a GP, it is possible to define a numerical relationship between variables. In the one dimensional case, given an input vector $\mathbf{x}$ and the corresponding output vector $\mathbf{y}$, it is possible to make predictions on a different input vector $\mathbf{x'}$, assuming a relationship specified by the kernel. Since GPs result in normal distribution, each input variable in $\mathbf{x'}$ will have a mean and a standard deviation as elements of the normal distribution, which could be interpreted as the regressed value and the error in the output variable.

The lightcurves take the form of flux vs. time, with several data points in different filters with associated error bands. However, due to irregularities in scanning the sky caused by factors such as seeing and weather, Most lightcurves have several regions of empty data in them (see Fig 1). It is necessary to interpolate between data points to be able to detect flares in the light curves and characterize them. GP regression has been traditionally used for interpolation of lightcurves and is useful in the context since they also provide an error band around the predicted outputs.

This project would use GP regression to interpolate light curves to obtain a predicted normal distribution for a time series from the beginning to the end of the light curve data at a cadence of 2 days. The mean prediction values would be used as a continuous function to fit the light curve while the standard deviation would provide the associated uncertainty. To account for the damped random walk nature of nuclear flares, a Matern-3/2 kernel would be used for the GP regression.[4]

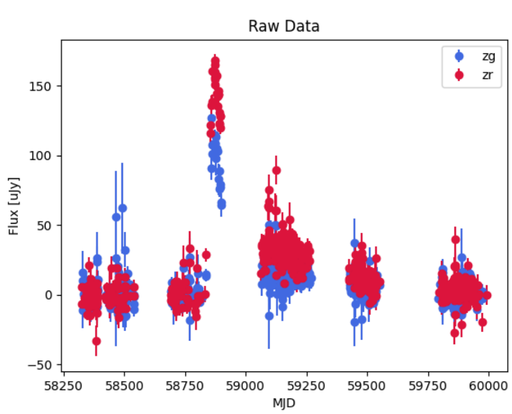

Figure 1. A sample raw light-curve after zero point adjustments. Annual observational periodicity could be seen through the gaps between data point clusters.

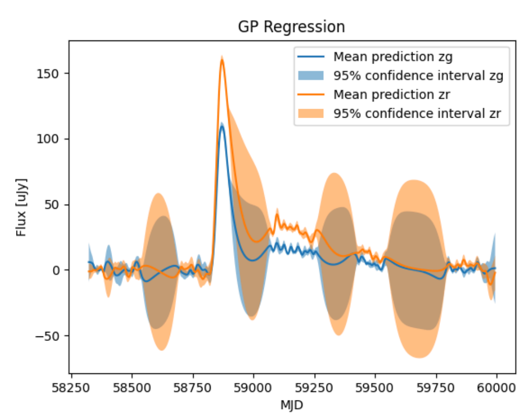

Figure 2. GP regression with the error bands for the light curve data in Figure 1. Regions with several data points have a lower uncertainty compared to those with no data points.

B. Exponentially smoothed moving average (ESMA) filter

Exponential smoothing is a technique used for smoothing time series data with the exponential window function.[8] It is calculated as a weighted average of the inputs with the weights exponentially decreasing over time. The simplest form takes the following formula:

$$s_t = \alpha x_t + (1-\alpha)s_{t-1} = s_{t-1} + \alpha(x_t - s_{t-1}).$$

Where x_t is the raw input at time series point t, s_t is the smoothed average, and α is a smoothing factor $0 \leq \alpha \leq 1$.

If x_t has a specific range, for instance, from -1 to +1, so does s_t. In this case, a contiguous set of observations of x_t 's with values above a certain threshold will smooth out s_t above the same threshold, after a certain time range of such observations, depending on the value of α. Thus, the attaining of s_t above the specified threshold can be used as a 'filter' to detect certain continuous patterns in the values of x_t. The smoothing ensures that it takes a certain minimum number of such observations x_t either to be detected or not detected by the filter.

The project would utilize ESMA filters to smooth out change in flux values over adjacent time series data points to detect contiguous increases and decreases in flux values, which can eventually be used to identify flares. The mean function, $y(t)$, and the standard deviations, $\sigma(t)$, obtained from the GP regression are used to calculate the raw input value for the filter:

$$raw(t) = \frac{y(t_i) - y(t_{i-1})}{\sqrt{\sigma(t_i)^2 + \sigma(t_{i-1})^2}}$$

The division by the uncertainty in flux change accounts for the uncertainty in GP regression. The raw value is then fed into a step function, described in table 1, to obtain an x_t value in the range of -1 to +1 which would be smoothed using the ESMA filter. Accordingly, a positive value of x_t indicates an increase in the flux and a negative value indicates a decrease in the flux. A higher absolute magnitude, $|x_t|$, is correlated with a steeper change in the flux and low uncertainty. The usage of a step function for mapping x_t to $[-1, 1]$ is mainly because of its simplicity.

The x_t values are then smoothed using the EMA formulae, with $\alpha = \frac{\Delta t}{T}$ with $T = 10$ days and $\Delta t = 2$ days since we chose a cadence of 2 days. A rise of smoothed value s_t above 0.8 would be flagged as an onset of a flare. If an onset is then followed by a decline of the smoothed value below -0.8, it is considered and detected as a flare.

Value of $raw(t)$	**Assigned value for** x_t
$0.2 < raw(t)$	+1
$0.06 < raw(t) \leq 0.2$	+0.5
$0.03 < raw(t) \leq 0.06$	+0.3
$-0.03 \leq raw(t) \leq -0.03$	0
$-0.06 \leq raw(t) < -0.03$	-0.3
$-0.2 \leq raw(t) < -0.06$	-0.5
$raw(t) < -0.2$	-1

Table 1. Outputs of the utilized step function to generate x_t.

C. Physical parameters of the flares.

The detected flares would be further studied to extract several physical parameters of them. The parameters are chosen from their usage in distinguishing TDE flares for the final phase of the project. The parameters included in the code are: peak flux value, *g-r* color at peak of the flare, time taken from half of peak to the peak value, time taken from peak to half peak value, and the rate of change of *g-r* color over measured for 5 time series data points after peak.

D. Approach

The first step would be to obtain an archival galaxy catalog for which time series data from ZTF, i.e. the light curves of the galaxies, can be obtained. The light curve data would then be parsed to obtain flux values and associated errors for the data points which are originally in magnitudes in two different optical filters: green and red.

The parsed light curves would then be fitted using GP regression and run through ESMA filters for detecting flares as outlined in previous sections. Then the flare physical parameters would be extracted from the detected flares and stored for further analysis. These flare parameters would then be studied alongside the several properties of the host galaxies to analyze and identify any correlations between the galaxy properties and the rate of flares. The physical parameters obtained for each flare would be used to distinguish between TDE flares and others.

IV. Progress

A. Previous studies on flare identification and TDE flare characterization.

The first week was spent on reading and understanding papers that focus on the identification of flares from lightcurves. Special attention was given to the way flares are characterized and defined in the papers and the numerical and statistical methods utilized to identify those. The papers also included those which focused on quasar variability and AGN flares.[2,4,6]

B. Code

The processing of light curves for the detection of flares was implemented through a Python program. The code was being developed and optimized from week 2 to week 7 with extensive experimentation on its several aspects around 4000 locally stored sample light curves. The code is available at:

B.1 Parsing the light curves.

The flux values were adjusted by deducting a weighted average of the data points before modified Julian date 58500, with the weights being chosen as $\frac{1}{(error)^2}$. This choice allows us to make zero point correction based on the left most data points and ensures that the color calculations do not have any negative values as inputs for the logarithms.

B.2. GP regression.

Further reading was done on gaussian processes and their implementation for interpolation of functions through papers and online resources. It was first implemented in the Python program using the module available in the `scikit-learn` Python library. However due to the latency in its performance, GP regression was re-implemented with `george` library, which made it approximately 2 times faster to make GP fits.

B.3. ESMA filter and flare parameters

These were implemented in the Python program as outlined in the previous section using basic numerical methods. The numerical figures associated with the filter, the smoothing window (T), threshold for detection of onset and offset of flares (+0.8 and -0.8), values chosen in the step function utilized to generate x_t , were all chosen based on experimentation with the locally stored samples. After identification of the flares, each flare is analyzed through a function of the Python program which could extract the flares' physical parameters.

B.4. Optimization of the code to be run on Big Data

All the previously mentioned parts of the code were developed between Week 2 and 6 and were run and tested on the locally stored samples. During week 7, the code was further optimized by removing unnecessary memory usage and debugging to prepare it to be used on a larger data set. In order to run the code faster, multiprocessing was implemented using Python's `ProcessPoolExecutor` module, which would help divide a larger dataset into smaller batches and run each batch on a different core on the computer. The code was also made user-friendly so that it could be run simply by specifying the required storage paths of the light curves. It is also made customizable to meet the different requirements of the user: the user can tune parameters, customize plots, save processed light curves as pickle files for future usage, adjust multiprocessing parameters (no. of cores and batch size).

C. Galaxy Catalog

To prepare the bigger data set for running the code to analyze and identify any physical correlations between flares, their rates and host galaxy properties, a galaxy catalog was created.

A list of galaxies with spectroscopic data available in Sloan Digital Sky Survey (SDSS) with redshift $z < 0.1$ was generated and was cross matched with PANSTARRS-DR2 to obtain their PANSTARRS IDs. These PANSTARRS IDs will be soon used to extract light curves from the ZTF. The galaxy catalog contains 412 225 galaxies which have spectroscopy in SDSS.

V. Future Work

In the upcoming weeks we plan to run the code on the large data set of ZTF light curves. Following any tuning or debugging of the code as needed, we will then analyze the data to obtain any important metrics, such as no. of flares, rate of finding flares, average peak values. The data would then be studied alongside the properties of the host galaxy, such as redshift and galaxy mass, with the anticipation of finding some correlations between the nature of the flares and the properties of host galaxies. For instance, we expect light curves with a single flare with short half-peak to peak time times, which are most probably TDEs, to occur in galaxies with mass less than around $10^{10}M_{\odot}$.[2] Such correlations obtained would help us better understand the nature of TDEs and their properties. If time permits, we also plan to further tune the parameters used in the process such as ESMA filter variables, the generator function for x_t.

VI. References

[1] Gezari et al., (https://ui.adsabs.harvard.edu/abs/2021ARA%26A..59...21G/abstract)

[2] Yao et al., Tidal Disruption Event Demographics with the Zwicky Transient Facility: Volumetric Rates, Luminosity Function, and Implications for the Local Black Hole Mass Function (https://arxiv.org/pdf/2303.06523.pdf)

[3] Charalampopoulos et al., AT 2020wey and the class of faint and fast tidal disruption events. (https://ui.adsabs.harvard.edu/abs/2023A%26A..673A..95C/abstract) (2023)

[4] Summer A. J. McLaughlin et al., Using Gaussian Processes to detect AGN flares (https://arxiv.org/pdf/2403.05354) (2024)

[5] Scikit-learn implementation of Gaussian process regression: https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html

[6] Graham et al., Understanding extreme quasar optical variability with CRTS - I. Major AGN flares (https://ui.adsabs.harvard.edu/abs/2017MNRAS.470.4112G/abstract) (2017)

[7] C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, (2006)

[8] Wikipedia page on exponential smoothing: (https://en.wikipedia.org/wiki/Exponential_smoothing#Double_exponential_smoothing_(Holt_linear))