Kepler's Sieve:
Learning Asteroid Orbits from Telescopic Observations

A dissertation presented

by

Michael S. Emanuel

to

Institute of Applied Computational Science

in partial fulfillment of the requirements

for the degree of

Master of Science

in the subject of

Data Science

Harvard University

Cambridge, Massachusetts

May 2020

Dissertation Advisors:
Professor Pavlos Protopapas
Professor Christopher H. Rycroft

Author:
Michael S. Emanuel

Kepler's Sieve:
Learning Asteroid Orbits from Telescopic Observations

Abstract

A novel method is presented to learn the orbits of asteroids from a large data set of telescopic observations. The problem is formulated as a search over the six dimensional space of Keplerian orbital elements. Candidate orbital elements are initialized by randomly choosing the six elements independently at random, either by sampling from a known asteroid or from uniformly distributed angles. A statistical model describes the distribution of angular distances between the directions of observed detections to the predicted direction from the observatory site to a body with these orbital elements. This model yields a log likelihood function, which attains large positive values when the candidate orbital elements are near to the elements of a real asteroid viewed many times in the data. The candidate elements and the parameters describing the mixture distribution are jointly optimized using gradient descent. Computations are performed quickly and efficiently on GPUs using the TensorFlow library.

The methodology of predicting the directions of telescopic detections is validated by demonstrating that out of approximately 5.69 million observations from the ZTF dataset, 3.75 million (65.71%) fall within 2.0 arc seconds of the predicted directions of known asteroids. The search process is validated in multiple stages. First, I demonstrate that given an initial guess equal to a perturbation applied to the elements of asteroids present in the data, the search process is able to successfully recover the true orbital elements to a high degree of precision. Next, I demonstrate that a random initialization is able to converge to elements matching some known asteroids to high precision. Finally, I use the search on observations that do not match any known asteroids. I present orbital elements for [5] new, previously unknown asteroids.

Exact number of new asteroids presented

All code for this project is publicly available on GitHub at github.com/memanuel/kepler-sieve.

Contents

Acknowledgments

I would like to thank my advisor, Pavlos Protopapas, for suggesting this topic and for his consistent support, advice and encouragement.

I would like to thank Chris Rycroft, my secondary advisor, for guiding me through a paper in Applied Math 225 in which I explored numerical integrators for solving solar system orbits.

I would like to thank Matt Holman and Matt Payne from the Center for Astrophics (CFA) for their advice on state of the art solar system integrators.

Most importantly, I would like to thank my wife Christie and my children Victor and Renée for their love and support. The Covid-19 crisis struck just as my work on this thesis kicked into high gear. It would not have been possible to complete it without extraordinary understanding and support from them.

Introduction

Determining the orbits of asteroids is one of the oldest problems in astronomy. Classical methods are based on taking multiple observations of the same body through a telescope. For an object that is large and bright enough, the human eye can ascertain the continuity of the motion, i.e. that the data are multiple sightings of the same object. Once enough sightings have been established, orbital elements can be solved using traditional numerical methods, such as a least squares fitting procedure that seeks elements to minimize the sum of squares error to all of the observations.

State of the art techniques for solving this problem are remarkably similar in spirit to the classical method. Indeed, the first interstellar object, 'Oumuamua, was discovered when astronomer Robert Weryk saw it in images captured by the Pan-STARRS1 telescope on Maui. [1] [2] More automated methods also exist. Still, these methods are based on a search in the space of the observable data attributes: the time of observation (MJD), the right ascension (RA) and declination (DEC). The apparent magnitude or brightness (MAG) is the third important observed quantity available for telescopic detections. Two observations made close together in time at two points in the sky very near to each other have a relatively high probability of belonging to the same object. Such a pair of observations in called a "tracklet." Today's automated approach to identifying new asteroids from telescopic data is based on performing a greedy search of the observed data to extend tracklets. Once a tracklet is identified, the algorithm attempts to extrapolate the path where future detections of this object might be. After enough detections are strung together, a fitting procedure is tried to determine the orbital elements.

This is a solid technique and I do not mean to cast aspersion on it. In this paper however I

[1] Wikipedia - Oumuamua

[2] NY Times - Astronomers Race to Study a Mystery Object from Outside the Solar System

propose a new method which I believe has some significant advantages. Rather than searching in the space of the data, i.e. (MJD, RA, REC, MAG), I propose to instead search over the six dimensional space of Keplerian orbital elements $(a, e, i, \Omega, \omega, f)$. Why should we complicate things by searching implicitly as it were on the space of possible orbits, rather than simplier and more direct method currently used? The main reason to switch is to avoid a combinatorial explosion.

If you limit your search to candidate tracklets where you detect the same object multiple times in a row, you are going to miss out any object that you detect only once or twice on a given night of observations. But if this same object were seen on multiple nights, posible separated over multiple days or longer, it becomes very costly to propose enough candidate tracklets to pick them up. Indeed you will soon face a combinatorial explosion in the number of possible tracklets. A simplified model of the number of tracklets might be that we have a data set containing N total observations, and we set a threshold τ in time and d in angular distance for how close a second observation must be to mark it as a candidate tracklet.

Here is a simple model showing the quadratic cost of enumerating candidate tracklets.
If you extend this further to tracks with 3 observations, the scaling gets even worse (cubic).
Let ρ be the average density of detections per day per degree of sky.
Let T be the number of days of observations in our data set.
Let τ be the threshold in days for 2 observations to be considered close enough in time to form a candidate tracklet.
Let d be the threshold distance in degrees for 2 observations to be considered close in the sky to form a candidate tracklet.
Let $A = 41,253$ be the number of square degrees in the sky. [3]
Let $N = T \cdot A \cdot \rho$ be the total number of detections in the data set.
Let $m = \tau \cdot \pi d^2 \cdot \rho$ be the average number of observations that will be close enough to each candidate starting point of a tracklet.
Let $NT_2 = \dfrac{N \cdot m}{2!} = \dfrac{\tau}{T} \cdot \dfrac{\pi d^2}{A} \cdot \dfrac{\rho^2}{2!}$ be the total number of candidate trackets of size 2.
Let $NT_k = \dfrac{N \cdot m^{k-1}}{k!} = \left(\dfrac{\tau}{T} \cdot \dfrac{\pi d^2}{A}\right)^{k-1} \cdot \dfrac{\rho^k}{k!}$ be the total number of candidate trackets of size k.
We can see the bad news right away. The number of tracklets of size k, NT_k, scales as ρ^k. And the

[3] Wikipedia - Square Degrees in the Sky

factors in the denominator don't bail us out. The number of possible ways m to extend a tracklet is going to be a large number well in excess of 1.

This is the principal motivation for searching in the space of orbital elements. While it's a large 6 dimensional space, it size is fixed. The cost of the search algorithm presented below scales linearly in the observation density ρ. The second major reason for searching in the space of orbital elements is that it permits the search algorithm to string together observations made far apart in time. This is a capability that eludes searches based on tracklets.

I summarize now the key steps in the search algorithm. The first step is to generate a set of candidate orbital elements. This is done with a very simple approach, one which can almost certainly be imporved on later: random initialization. For four of the orbital elements, a, e, i, and Ω, one of the 780,000 catalogued asteroids is selected at random. Its orbital elements are used to populate these four. The remaining two orbital elements M (mean anomaly) and ω (argument of periapsis) are modeled to be distributed uniformly at random on the circle $[0, 2\pi)$. These are then converted to the representation using $(a, e, i, \Omega, \omega, f)$.

Once the candidate elements have been initialized, they are integrated numerically using the `rebound` library. This is considered to be the gold standard of their true orbits. This initial integration is then used to filter the data set of ZTF observations to a subset that are relevant for searching for orbits. A routine computes the right ascension (RA) and declination (DEC) that an observer at a given observatory site on earth would have seen light leaving an object with the candidate elements at a given observation time (MJD). This quantity is computed at each unique observation time in the ZTF data set. The angular distance between the predicted and observed direction is computed. A threshold (2.0 degrees) is applied, and all ZTF observations falling within this threshold are cached in memory of the search class.

During the main body of the search process, the elements will be adjusted by a small amount in each training round. These perturbed elements will have their orbits evaluated using the Kepler 2 body model. An implementation is performed on the GPU using TensorFlow that is fast and differentiable. The ground truth orbit is used to provide an adjustment term so that the predicted orbits will match the true orbits exactly when the perturbation is zero. The predicted orbit can therefore be considered to be a linearization of the true orbits based on the Kepler model.

The objective of the optimization function is based on the log likelihood of a statistical model for the distribution of distances between predicted and observed directions. A lemma will demonstrate that for directions uniformly distributed on the sphere, the squared distance over the threshold distance would be uniformly distributed on the interval $[0, 1]$. A mixture model is formulated, where the distance between every predicted and observed direction is modeled as a mixture of hits and misses. The misses are distributed uniformly on $[0, 1]$. The hits are distributed as a truncated exponential distribution. The decay parameter λ of this exponential process is associated with a resolution paramater R. This model is equivalent to assuming that some fraction h (for hits) of the detections are due to a real body with the candidate elements, and that the results of the detection will be normally distributed with a precision parameter equal to the resolution. During the search process, the threshold parameter is also updated. This dynamic threshold should not be confused with the original threshold of 2.0 degrees used to build the filtered training data.

The optimization process jointly optimizes the candidate orbital elements and three parameters in the mixture model: the assumed number of hits, the resolution R, and the threshold. Intuitively, we want the model to gradually tighten its focus, and adjust the orbital elements so they hit as many observations as closely as possible. Early on, the optimization will probably just try to get close to the central tendency of the data set. If the initialization was good, it will gradually tighten in the resolution and threshold parameters. The gradients will encourage the model to adjust the candidate orbital elements so that some of the observations, the ones it is implicitly modeling as hits, will be very close what is predicted by the candidate elements. The observations modeled as highly probable misses will hardly contribute to the gradients of the candidate elements.

In practice, the optimization is actually carried out in alternating stages. In odd numbered stages, only the resolution parameters are tuned at a higher learning rate; in even numbered stages, both the resolution and orbital elements are adjusted together at a slower learning rate. There are some additional subtleties where the actual optimization function during the training of the mixture parameters has a term to encourage the model to shrink the resolution and threshold parameters. These will be discussed at greater length below.

As much as possible, I have sought to validate individual components of these calculations in isolation. My numerical integration of the planets is validated against results from NASA JPL

(Jet Propulsion Library) using the superb Horizons system [4] I separately validated the numerical integration of the first 20 asteroids.

GET EXACT NUMBER

The notion of a direction in space from an observer on earth is typically reported in telescopic data using a right ascension and declination. While these are convenient and standard for reporting observed data, they are not well suited to the approach taken here. All directions are represented internally in this project as a unit vector $\mathbf{u} = (u_x, u_y, u_z)$ in the Barycentric Eliptic Plane. These calculations were validated in isolation by querying the Horizons system for both the positions and directions to known asteroids.

The end to end calculation of a direction from orbital elements was verified indirectly as follows. I integrated the trajectories of all 780,000 known asteroids using a collection of orbital elements downloaded from JPL. I then computed the nearest asteroid number to each ZTF asteroid, and the distance between the predicted direction and observed direction. I reviewed the statistical distribution of these distances. I observed that out of approximately 5.69 million observations from the ZTF dataset, 3.75 million (65.71%) fall within 2.0 arc seconds of the predicted directions of known asteroids. I took this as overwhelming evidence that these calculations were accurate.

To put this degree of precision in context, Pavlos has estimated that 1.0 arc second is a good back of the envelope estimate of the precision with which a modern telescope can determine direeection of an observation. If you were to use an approximation that observations were made at Earth's geocenter (i.e. you did not account for location of the observatorory on Earth's surface) you would already be making errors on the order of 3 arc seconds. If you were to perform your calculations using the sun's location as your coordinate original rather than the solar system barycenter, you would make errors larger than 1.0 arc second. I know because I made both of these errors in earlier iterations before squeezing them out!

I tested the capabilities of the search process with an increasingly demanding set of search tasks. The first block of search tasks involved recovering the elements of known asteroids. I took a batch of 64 asteroids that appeared most frequently in the ZTF data set. These asteroids were

[4] NASA Horizons
I cannot say enough good things about Horizons. If you want an external "gold standard" of where an object in the solar system was or will be and a friendly user interface, Horizons has you covered.

represented between 160 and 200 times in the data, where hits here are counted at a threshold of 2.0 arc seconds as before. Here is a summary of the tests I ran:

- Initialize search with correct orbital elements, but resolution $R = 0.5°$ and threshold $\tau = 2.0°$. All 64 elements were recovered to **???**
- Initialize search with small perturbation applied to orbital elements; a by 1.0%, e by 0.25%, i by 0.05°, remaining angles f, Ω and ω by 0.25°. 37 of 64 elements were recovered to **???**.
- Initialize search with large perturbation applied to orbital elements; a by 5.0%, e by 1.0%, i by 0.25°, remaining angles f, Ω and ω by 1.0°. 11 of 64 elements were recoverd to **???**. In some cases, a different (but correct) set of orbital elements was obtained; the perturbation was so large the search found a different asteroid.
- Initialize a search with **randomly initiialized** orbital elements. Search against the subset of ZTF observations within 2.0 arc seconds of a known asteroid. This search converged on one set of orbital elements matching a real asteroid.

The last last test was significantly more demanding in that it did not rely on known orbital elements.

The work presented above can be seen as a way to indpendently validate a subset of the known asteroid catalogue. It can efficiently associate a large number of telescope observations with known asteroids, which could in turn be used to further investigate those asteroids. Analysis might include refining their estimated orbital elements, fitting the $H - G$ model of brightness, or identifying some of them for further investigation if they meet criteria of interest, e.g. orbits that will approach near to earth in the future.

The main thrust of this work however is not on refining the existing asteroid catalog, it is finding new asteroids. The final search I ran was against the subset of ZTF observations that did not match any of the known asteroids. Random initializations for orbital elements were tried. Most of these initializations fail to converge on elements with enough hits to match real asteroids in the data, but a small number do successfully converge. So far I have identified 10 asteroids with 8 or more hits. I have verified that none of the orbital elements modeled for these asteroids appear in the catalogue of known asteroids I obtained from JPL. I have also done an ad-hoc review of the ZTF records to ensure that they are plausibly belonging to the same object. I believe that these represent new and unkown asteroids, and plan to submit them to the Minor Planet Center for possible classification. **UPDATE WITH REAL NUMBERS**

The ultimate goal of this project is not to simply perform a one time search of a dataset to identify some new asteroids. The goal is rather to create a tool that will be of enduring use to astronomy community for solving the problem of searching for new asteroids given large volumes of telescopic data. To that end, I plan to consult with Matt Holman and his colleagues at the Minor Planet Center to see what refinements and improvements would be required to upgrade this from a tool I can use to one that is of wider use to the astronomy community.

Chapter 1

Integrating the Solar System

1.1 Introduction

Block Quotations (quotation and quote environments) are supposed to be single-spaced with each entry, and double-spaced between. The class file does this automatically. For example:

> Dummy quote. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

> Dummy quotation. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Table 1.1: *Table heading goes on top of the table*

Tables	should
Be	double
spaced	unless
they are	long
This	table
is	getting
long	
so	I
manually	
set	it
to	single
spacing using	

Table 1.2: *Use consistent format for captions*

Table	should	be	placed
within	text,	as	close
to	its first mention		
as	possible.	Not at the end	
of a chapter	or dissertation		

1.2 Motivating Example

Table 1.1 shows stuff. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Table 1.2 shows stuff also.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.3 Introduction

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. $\sin^2(\alpha) + \cos^2(\beta) = 1$. If you read this text, you will get no information $E = mc^2$. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$. This text should contain all letters of the alphabet and it should be written in of the original language. $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$. There is no need for special content, but the length of words should match the language. $a\sqrt[n]{b} = \sqrt[n]{a^n b}$.

1.4 Potential outcomes framework

Hello, here is some text without a meaning. $\mathrm{d}\Omega = \sin\vartheta \mathrm{d}\vartheta \mathrm{d}\varphi$. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. $\sin^2(\alpha) + \cos^2(\beta) = 1$. This text should contain all letters of the alphabet and it should be written in of the original language $E = mc^2$. There is no need for special content, but the length of words should match the language. $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$. [1][2]

1.5 Conclusion

I conclude that:

- First item in a list
- Second item in a list
- Third item in a list
- Fourth item in a list
- Fifth item in a list

[1] Footnotes are single-spaced. Hello, here is some text without a meaning. $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$. This text should show what a printed text will look like at this place. $a\sqrt[n]{b} = \sqrt[n]{a^n b}$. If you read this text, you will get no information. $\mathrm{d}\Omega = \sin\vartheta \mathrm{d}\vartheta \mathrm{d}\varphi$. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. $\sin^2(\alpha) + \cos^2(\beta) = 1$.

[2] Space between foonotes is doublespaced. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. $\sin^2(\alpha) + \cos^2(\beta) = 1$. If you read this text, you will get no information $E = mc^2$. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$. This text should contain all letters of the alphabet and it should be written in of the original language. $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$. There is no need for special content, but the length of words should match the language. $a\sqrt[n]{b} = \sqrt[n]{a^n b}$.

Chapter 2

Predicting Directions from Positions

Chapter 3

Searching for Asteroids

3.1 Introduction

Some people just cite papers in introductions for no reason. Anderson and Rubin (1949); Pearson (1901); Spearman (1904).

3.2 Setup

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. $\sin^2(\alpha) + \cos^2(\beta) = 1$. If you read this text, you will get no information $E = mc^2$. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$. This text should contain all letters of the alphabet and it should be written in of the original language. $\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}}$. There is no need for special content, but the length of words should match the language. $a\sqrt[n]{b} = \sqrt[n]{a^n b}$. See Figure 3.1 for illustration.

3.3 Conclusion

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no

```
#include <iostream>
int main(int argc, char** argv) {
  std::cout << "Hello World." << std::endl;
  return 0;
}
```

Figure 3.1: *Captions for figures go at the bottom of the figure.*

information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

References

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, **20** (1), 46–63.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2** (11), 559–572.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, **15** (2), 201–292.