Visualizing the Chaos of the Sitnikov Problem

Chris Wang, MATH181

May 8, 2024

Contextualization

More than 300 years ago, the motions of celestial bodies in the sky were elucidated by physicists such as Johannes Kepler and Isaac Newton. Understanding how two bodies gravitationally interacted became a straightforward task and has, in the lens of classical mechanics, become an exceptionally well understood system.

Extending this ideal to three gravitationally interacting bodies, though, and literal chaos arises. The three body problem, as it came to be known, is one of the quintessential examples of a chaotic dynamical system. In a nutshell, chaotic systems satisfy the following conditions [4]:

1. **Aperiodic long-term behavior.** Given a small range of initial conditions, we see that the three body problem admits trajectories of all sorts, including ones that do not settle to fixed points or end up in (quasi)periodic orbits.

2. **Deterministic.** The system's chaos arises from its innate nonlinearity and not from random driving forces. The three body problem satisfies this: assuming three point masses only and assigning initial data to the system with no additional interference, we still get chaotic behavior.

3. **Sensitive dependence on initial conditions.** Trajectories initially arbitrarily close together eventually separate, and their separation is exponential. Technically speaking, this means the system has a positive maximal Lyapunov exponent. We will see this in action when we visualize trajectories later.

In this document, I hope to shed some light on how we can visualize the chaos of the three body problem. In particular, I am visualizing a special case of the problem, the *Sitnikov problem*. This system assumes two identically massive "parent" stars orbiting each other in ellipses with a third planet that moves normal to the ecliptic plane. The planet is assumed to cross through the barycenter of the stars and to also be significantly less massive than the parent stars, allowing us to effectively neglect the planet's mass.

My goal with this project is to show how the nonlinearity of the Sitnikov problem arises, visualize trajectories in the system, and provide intuition to the chaos.

Mathematical Basis

Before getting into the mathematics, I've included a figure below to visualize the Sitnikov problem:

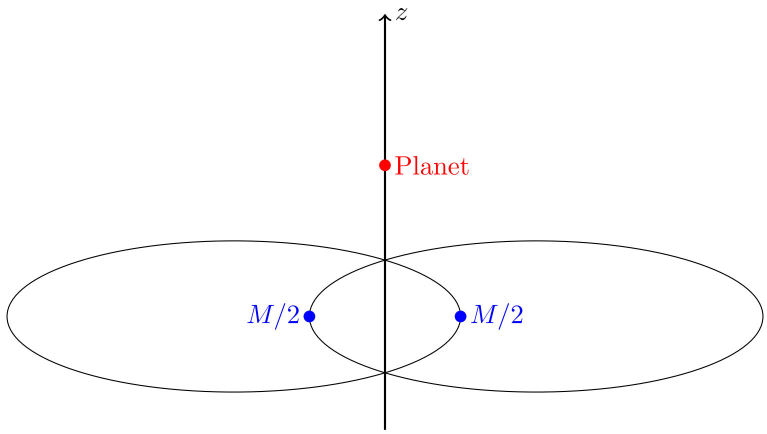

Figure 1: An elementary visualization of the Sitnikov problem. The planet starts on the z axis and is gravitationally attracted to the two parent stars, each with mass $M/2$. The reasoning for why I have let each star have mass $M/2$ becomes apparent soon.

We can now begin constructing the mathematics behind the problem. First, we make note of a few important parameters. We will denote $z = z(t)$ as the planet's distance above the ecliptic and $\dot{z} = \dot{z}(t)$ as the velocity of the planet. Thankfully, since the planet's motion is restricted to one dimension here, we need not worry about vector quantities. We then denote $r^2 = r^2(t) = x^2 + y^2 + z^2$ as the distance from the planet to the stars and $\rho^2 = \rho^2(t) = x^2 + y^2$ as the "ecliptic distance".

I now borrow an essential result from classical mechanics: the Euler-Lagrange equations. Given some Lagrangian L, which is related to the total energy of the system, the equations of motion for the system is essentially given by

$$\frac{\partial L}{\partial z} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{z}} = 0.$$

The Lagrangian itself is defined as the difference between the kinetic energy and potential energy of the system. In this case, the kinetic energy is simply $\frac{1}{2}m\dot{z}^2$, while the potential energy is just $-\frac{GMm}{r(t)^2}$. Hence,

$$L = \frac{1}{2}m\,\dot{z}^2 + \frac{GMm}{r(t)^2} = \frac{1}{2}m\,\dot{z}^2 + \frac{GMm}{(\rho^2 + z^2)^{1/2}}.$$

We take some derivatives:

$$\frac{\partial L}{\partial z} = -\frac{GMmz}{(\rho^2 + z^2)^{3/2}}, \quad \frac{\partial L}{\partial \dot{z}} = m\,\dot{z} \quad \implies \quad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{z}} = m\,\ddot{z} = -\frac{GMmz}{(\rho^2 + z^2)^{3/2}}.$$

Canceling out the m on both sides representing the planet's (practically negligible) mass, we have the final equation of motion, which is evidently nonlinear:

$$\ddot{z} = -\frac{GMz}{(\rho^2 + z^2)^{3/2}}. \tag{1}$$

By our construction, $\rho(t)$ is periodic, since the parent stars' orbits are periodic. Therefore, we can model the function as a Fourier series perturbatively, where we assume the eccentricity is much less than 1. As such, we can make the following assumption:

$$\rho(t) \simeq a\left(1 - \varepsilon \cos t\right), \text{ where } a \text{ is the semi-major axis and } \varepsilon \text{ the perturbation.} \tag{2}$$

From here, we can get a system of coupled first order differential equations by writing $\zeta = \dot{z}$, so that equation 1 becomes

$$\begin{aligned} \dot{z} &= \zeta \\ \dot{\zeta} &= -\frac{GMz}{\left(\rho^2 + z^2\right)^{3/2}} \end{aligned} \tag{3}$$

Using our form for $\rho(t)$ as in equation 2 as a substitution for ρ in equation 3, then performing half a million binomial approximations, we can write $\dot{\zeta}$ as

$$\dot{\zeta} \simeq -\frac{GMz}{\left(a^2 + z^2\right)^{3/2}} - \frac{3a^2 GMz\varepsilon \cos t}{\left(a^2 + z^2\right)^{5/2}}.$$

Evidently, this differential equation is very messy, and there are no known closed-form solutions.

The Unperturbed Case

In the special case where both stars have circular orbits, we can take $\varepsilon = 0$. As a result, we can rewrite equation 1 as

$$\ddot{z} + \frac{GMz}{\left(a^2 + z^2\right)^{3/2}} = 0 \tag{4}$$

While this still looks pretty bad due to the z in the denominator, we can just multiply everything by $\dot{z}$ and show that the differential equation is now separable:

$$\begin{aligned} \dot{z}\ddot{z} + \frac{GMz\dot{z}}{\left(a^2 + z^2\right)^{3/2}} &= 0 \\ \implies \quad \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\dot{z}^2 - \frac{GM}{\left(a^2 + z^2\right)^{1/2}}\right) &= 0 \\ \implies \quad \frac{1}{2}\dot{z}^2 - \frac{GM}{\left(a^2 + z^2\right)^{1/2}} &= \kappa. \end{aligned}$$

We could then integrate and get an exact equation of motion for the planet.

The Perturbed Case

Of course, what we are interested in is the perturbed case, where $\varepsilon > 0$. Now, the equations of motion are not clear, and to see where the chaos arises, we instead look at a phase space representation of the system. Using a polar plot and letting each point on the plot correspond to the moment when $z = 0$, we can plot $|dz/dt|$ as the "radius" of a point and the relative position of the stars as the "phase", since we have implicitly normalized the period of each star's orbit to be 2π. A phase of zero corresponds to when the stars are closest to the barycenter, since $\rho(0) = a(1 - \varepsilon)$. Intuitively, then, we might guess that there are certain regions of the plot (as initial conditions) that correspond to the planet escaping the star system. Some basic back-of-the-envelope calculations show that when $|dz/dt| > \sqrt{2}$ [1], the planet is bound to fly off to infinity. Hence, the region must be bound by a circle of radius $\sqrt{2}$.

The actual shape of the region, termed D_0 in the papers I read about the subject, can be determined through advanced measure theory and topology [3] [2] [1]. That is beyond the scope of my project; instead, I focused on gaining intuition for how the region behaved under a transformation ϕ, which mapped each point in D_0 to another point in phase space representing $|dz/dt|$ and the phase of the stars when the planet crossed the barycenter again.

Plots of the Trajectories

To understand where chaos might arise, particularly the sensitivity to initial conditions, I wrote a program in Python to plot planet trajectories with similar initial conditions. I chose a segment of phase space constant in argument (I arbitrarily chose $\pi/4$) and varying slightly in initial speed, from 1.4 to 1.4001. I also chose the perturbation value to be $\varepsilon = 0.25$, at least for Figures 2 and 3. This gave me test points close enough to the edge of D_0 that any sensitivities would be exaggerated. A sample plot is shown below, where I picked 127 points between $v_0 = 1.4$ and $v_0 = 1.4001$ to illustrate how even such fine spacings can lead to divergent trajectories, exhibiting sensitivity to initial conditions.

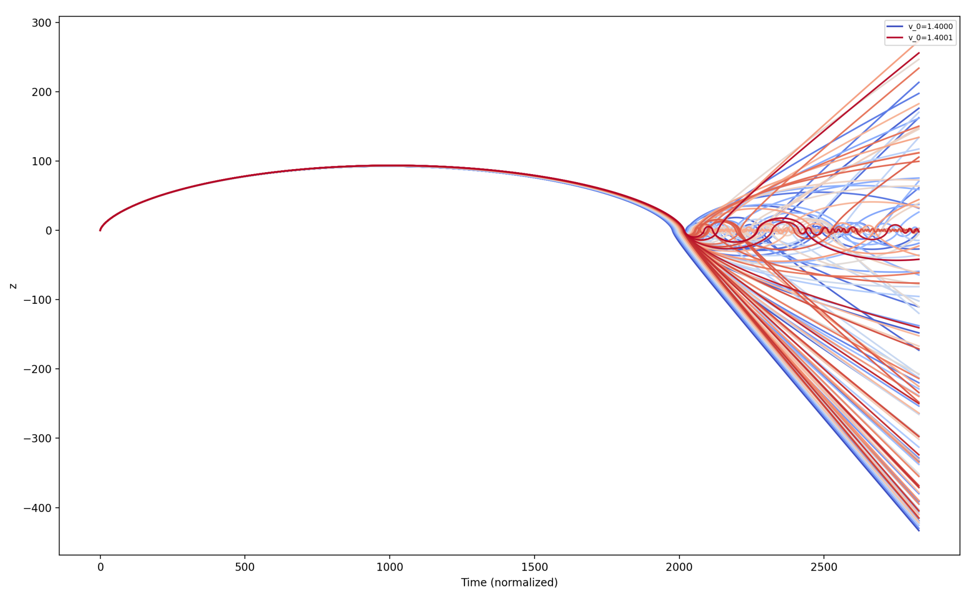

Figure 2: Solving the Sitnikov differential equation for trajectories with v_0 ranging from 1.4 to 1.4001 visualizes the system's sensitivity to initial conditions. The plot is done through lines 69-95 of `sitnikov.py`.

I also attempted to plot a similar figure in the reverse-time direction, however, for the specific range of v_0

[1]This can be calculated using conservation of energy and assuming the planet has zero velocity at infinity.

values chosen above, the resulting plot was kind of boring (all trajectories were initially divergent). What this means is that the set of initial conditions I plotted in Figure 2 were generated by the stars "capturing" the planet. A more interesting example is shown where $v_0 \in [1.34, 1.3401]$:

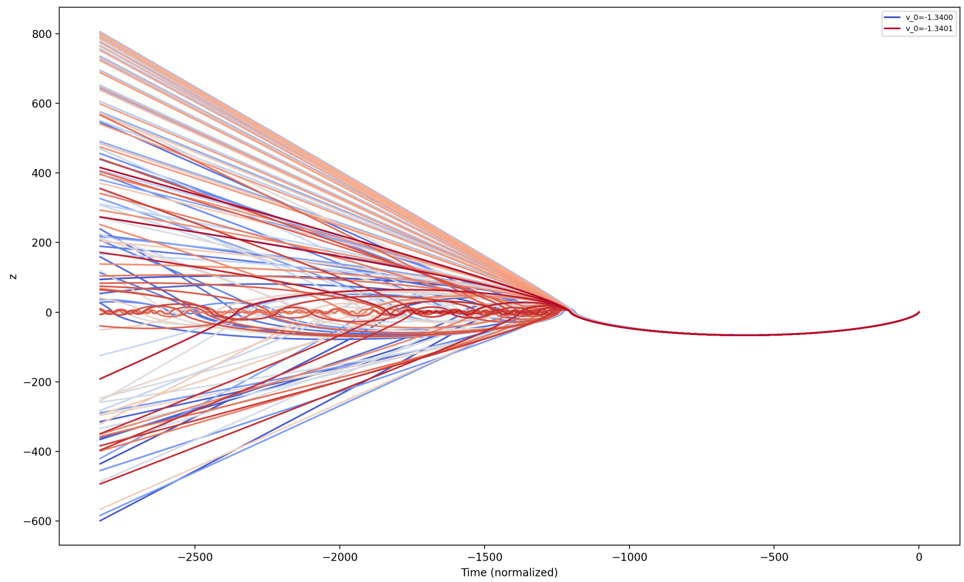

Figure 3: A reversed time plot for the Sitnikov system, showing trajectories that converged on the initial condition set such that $v_0 \in [1.34, 1.3401]$ and having phase $\pi/4$.

One can still notice that there's a consierable period where trajectories converged in Figure 3, but upon further experimentation, I found that solving the time-reversed Sitnikov to yield v_0 lower than $v_0 = 1.34$ correlated with trajectories that converged closer to $t = 0$.

The Strange Attractor

With these plots, I now wanted to investigate what the region $\phi(D_0)$ would look like. At least, I wanted a sense of how a set of neighboring trajectories in D_0 were modified by ϕ. The idea here is to show that we have a strange attractor in phase space, a hallmark of chaotic systems. In particular, I want to visualize how the map ϕ stretches and folds the set of initial conditions. The initial conditions I fed in assumed $\varepsilon = 0.1, v_0 \in [1.4, 1.401]$, and the initial phase of the stars to be $\delta = \pi/4$. I chose this set since it lies close to the boundary ∂D_0, which would give a better visualization of the horseshoe map.

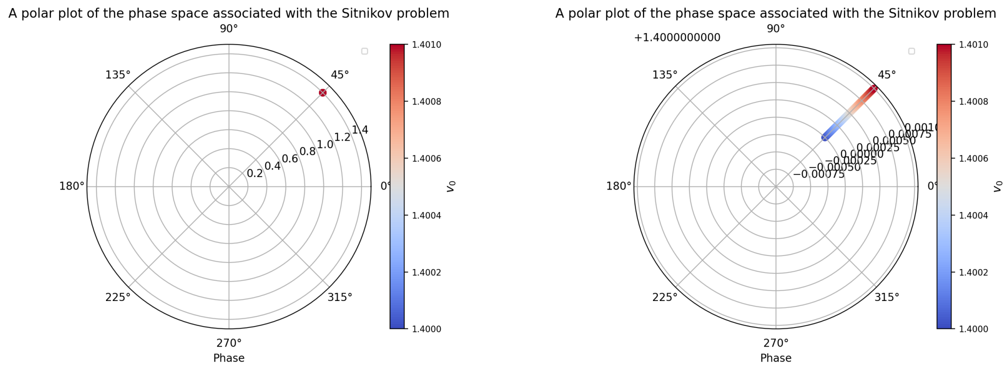

Figure 4: Side by side plots of the initial set. The plots are the same except the right plot has been enlarged to show the spectrum of initial velocities. Lines 126-154 of `sitnikov.py` yield this plot.

Using these initial conditions, I ran my program. The resulting phase space plot is shown below.

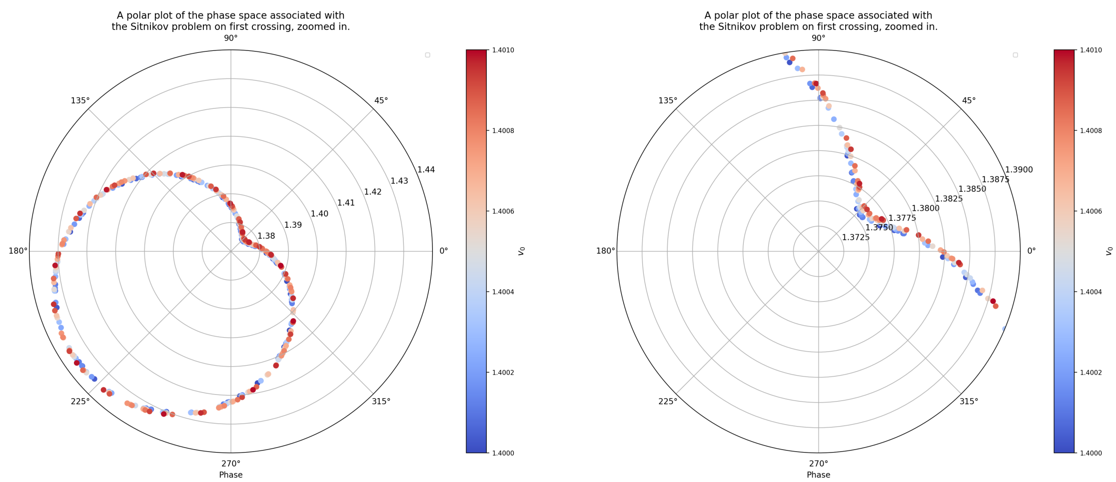

Figure 5: The set of initial conditions in Figure 4 after running the Sitnikov program. Note that both plots are zoomed in and the origin does not start at $r = 0$. If we denote the set in Figure 4 as γ, then the resulting set here is interpreted as $\phi(\gamma)$.

With this visualization, we can see that the initial set γ is stretched and looped around itself, traversing all phases. The bottom panel visualizes the finer structure of $\phi(\gamma)$. Observe how the points corresponding to higher v_0 have a larger radius at the same phase, in accordance with Moser's illustration of $\phi(\gamma)$.[3]

Intuitively, we can make sense of this. Upon close inspection of Figure 2, around normalized time $t \sim 2000$, the closely ordered trajectories return to barycenter at slightly different times, corresponding to slightly different phases in the phase space plot. Physics also tells us that, under oscillator-esque circumstances such as this, trajectories with higher v_0 also return with higher velocity, as evidenced by their comparatively larger radii in Figure 6. From this, then, the stretching and looping in phase space becomes apparent.

An interesting observation is that the resulting contour $\phi(\gamma)$ strongly resembles a cardioid. Its orientation is determined by the initial phase δ, its radius determined by v_0, and its distortion seems to be related to ε in some way. In summary, $\phi : \gamma \mapsto \phi(\gamma)$, as shown:

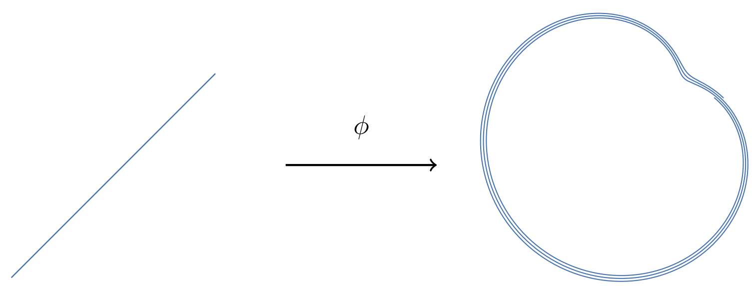

Figure 6: An illustration showing how ϕ stretches and folds γ. The top right point in γ maps to the endpoint with larger radius in $\phi(\gamma)$. Illustration done in Desmos.

A similar case can be made for the time-reversed set $\phi^{-1}(\gamma)$. Considering Figure 3, we see that trajectories eventually matching the initial conditions for γ do converge at normalized time $t \sim -1200$, which is also where they last cross the barycenter. Plotting the speeds and phases there should yield a similar plot as in Figure 5.

From here, we see that if we consider γ such that its endpoint lies on ∂D_0, even making γ an arbitrarily small segment means $\phi(\gamma)$ can be spread out over arbitrarily many periods of the star system, thus distorting γ into a curve that loops infinitely within $\phi(D_0)$. This is to say that given a small enough ε, there always exists an integer corresponding to the number of "Sitnikov years" a planet takes to return to the barycenter (this is essentially a restatement of Moser's theorem).[3]

Remarks

I note that in all the figures I've pasted, I only changed the v_0 value while keeping the initial phase δ constant. An area of further computational exploration would be to vary v_0 across a wider range and vary $\delta \in [0, 2\pi]$ in order to construct $\phi(D_0)$ completely. Mathematically, it wil also be interesting to see how the cardioid shape arises. It may be due to how $\rho(t)$ is constructed, but further investigation is needed.

References

[1] T. Kovacs and B. Erdi. Transient chaos in the sitnikov problem. *Celestial Mechanics and Dynamical Astronomy*, 2009.

[2] Jie Liu and Yi-Sui Sun. On the sitnikov problem. *Celestial Mechanics and Dynamical Astronomy*, 1990.

[3] Jugen Moser. *Stable and random motions in dynamical systems.* Princeton University Press, 1973.

[4] Steven Strogatz. *Nonlinear Dynamics and Chaos.* CRC press, 2 edition, 2015.