DMU Final Project

Jesse Greaves, Andrew Palski, Kevin Yevak

I. Introduction and Background

Cislunar space is an exciting proving ground for a myriad of future missions, from human habitation to autonomous exploration, due to the region's suitability as a staging point for deep space exploration. The concept of utilizing cislunar space as a waypoint for deep space exploration has been adopted by national space agencies and commercial companies alike. In particular, the multinational Artemis program is an exemplary strategy for cislunar development. The Artemis program includes plans for surface exploration, long term habitation, and technological demonstrations in the lunar sphere of influence [1]. One area of technological development which would greatly facilitate future lunar missions, and missions farther from Earth in general, are autonomous Guidance, Navigation, and Control (GNC) capabilities. As a whole, automating GNC systems for cislunar formation flight could reduce costs, enhance safety, and potentially enable novel operations. Specifically, autonomous GNC architectures would be exceptionally beneficial to the Lunar Gateway which is entirely dependent on formation flying and proximity operations for both its construction and operation.

Fully autonomous formation flying requires accurate real-time navigation. Ideally, this would be addressed by solving the co-estimation problem only utilizing onboard sensors. The co-estimation problem is when a chaser vehicle attempts to simultaneously estimate its own state as well as the state of another target vehicle. This is equivalent to solving for both the relative and absolute states of the system. The relative sensor co-estimation problem for cislunar spacecraft was first solved using cross-linked range measurements by LiASION [2]. Then additional studies expanded those results to optical measurements [3]. Optical sensors not only provide improved observability, but they can also leverage techniques from bearing-only navigation literature. One such topic of optical navigation literature has shown that optics are sufficient for autonomous single craft navigation in cislunar space [4, 5], and therefore it is feasible to perform end-to-end autonomous cislunar navigation using a single senor system. Altogether, optical measurements are an attractive solution to the cislunar co-estimation problem since they require minimal power, weight, and cost yet still enable autonomous GNC for all phase of flight.

While optical sensors are an appealing approach to autonomous navigation, it cannot be neglected that they lack range information. The missing range content presents a significant challenge to general co-estimation, since relative range is unobservable in linear systems [6, 7]. Fortunately, in non-linear systems higher order dynamical moments can account for range to make it observable; hence the success of cislunar co-estimation [2, 3]. Unfortunately, the range estimate is often only weakly observable and highly dependent on the nominal trajectory. Additionally, the same non-linearities that produce observability also cause issues for linearized filters. The linearization assumptions are also strained by the inflated uncertainty profiles associated with the co-estimation problem, further reducing filtering performance. Thus, it becomes increasingly vital to obtain additional state information to alleviate non-linearities, improve observability, and ensure accurate navigation.

Bearing-only navigation techniques to obtain relative range information have been well studied in the Low Earth Orbit due to vast experience with the International Space Station [8, 9]. Recently, similar work examined this type of information gathering in cislunar space with very promising results [10]. Crucially, all the previous work assumed a well-known target state so only a relative state needed to be estimated. Additionally, the prior cislunar work only set a boundary condition on the observability angle as a way of enforcing information gain. These previous techniques set the foundation for new guidance policies which seek to maximize range information in the cislunar co-estimation problem given relative optical measurements, which is the goal of this work.

II. Problem Statement

A. System Definition

The sub-state of a single vehicle contains the position and velocity of that vehicle, as denoted by Equation 1. In this paper we are interested in the cislunar co-estimation problem where both the chaser and target spacecraft are simultaneously estimated. Therefore, the state of the system contains the sub-state of each vehicle as given by Equation 2, with independent state dynamics as denoted by Equation 3.

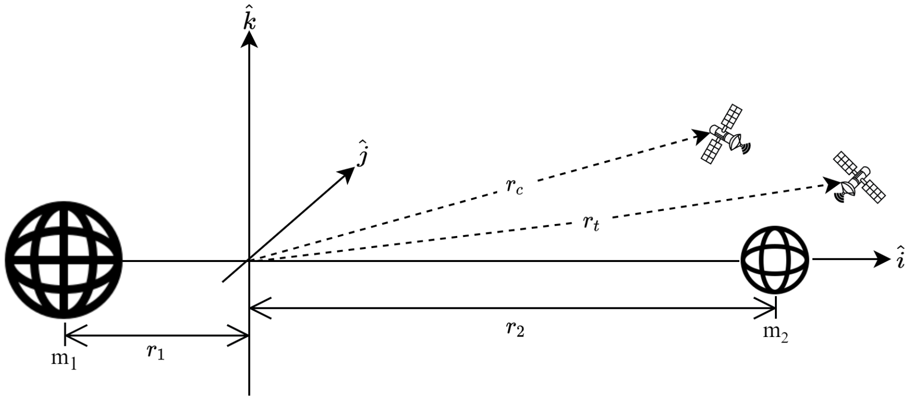

Fig. 1 Diagram of CR3BP in rotating frame with chaser and target.

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{r}^T & \boldsymbol{v}^T \end{bmatrix}^T \tag{1}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_c^T & \boldsymbol{x}_t^T \end{bmatrix}^T \tag{2}$$

$$\dot{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{f}(\boldsymbol{x}_c)^T & \boldsymbol{f}(\boldsymbol{x}_t)^T \end{bmatrix}^T \tag{3}$$

To approximate cislunar motion both vehicles are governed by the Circular Restricted 3 Body Problem (CR3BP). Figure 1 contains a diagram of the CR3BP in the rotating frame, and Equation 4 gives the dimensional equations of motion in this frame as well. Note that $\boldsymbol{r}_1$, m_1 and $\boldsymbol{r}_2$, m_2 are the position and mass of the primary and secondary bodies respectively. Spacecraft positions relative to the massive bodies are shortened to $\boldsymbol{\rho}_1 = \boldsymbol{r} - \boldsymbol{r}_1$ and $\boldsymbol{\rho}_2 = \boldsymbol{r} - \boldsymbol{r}_2$. Lastly, the cross product is replaced by its matrix representation such that $[\tilde{\boldsymbol{n}}]\boldsymbol{r} = \boldsymbol{n} \times \boldsymbol{r}$.

$$\begin{gathered} \boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{v}^T & \boldsymbol{a}(\boldsymbol{x})^T \end{bmatrix}^T \\ \boldsymbol{a}(\boldsymbol{x}) = -\frac{Gm_1\boldsymbol{\rho}_1}{||\boldsymbol{\rho}_1||^3} - \frac{Gm_2\boldsymbol{\rho}_2}{||\boldsymbol{\rho}_2||^3} - 2[\tilde{\boldsymbol{n}}]\boldsymbol{v} - [\tilde{\boldsymbol{n}}]([\tilde{\boldsymbol{n}}]\boldsymbol{r}) \\ \boldsymbol{n} = \begin{bmatrix} 0 & 0 & \sqrt{\frac{G(m_1+m_2)}{||\boldsymbol{r}_2-\boldsymbol{r}_1||^3}} \end{bmatrix}^T \end{gathered} \tag{4}$$

The State Transition Matrix (STM) is used to propagate state deviations in a linearized framework and is a function of the Jacobian of the dynamics taken with respect to the state. The notation $\boldsymbol{\Phi}(t_i, t_0)$ is the STM from time t_0 to time t_i, and Equations 5-(6) define the STM. Note that the STM is a block diagonal because the two spacecraft are independent of each other, and thus their dynamics are not correlated.

$$\boldsymbol{\Phi}(t_0, t_0) = \boldsymbol{I}_{12} \tag{5}$$

$$\begin{gathered} \dot{\boldsymbol{\Phi}}(t_i) = \begin{bmatrix} \boldsymbol{A}(\boldsymbol{r}_c) & \boldsymbol{0}_6 \\ \boldsymbol{0}_6 & \boldsymbol{A}(\boldsymbol{r}_t) \end{bmatrix} \boldsymbol{\Phi}(t_i, t_0) \\ \boldsymbol{A}(\boldsymbol{r}) = \begin{bmatrix} \boldsymbol{0}_3 & \boldsymbol{I}_3 \\ \boldsymbol{A}_{2,1}(\boldsymbol{r}) & -2[\tilde{\boldsymbol{n}}] \end{bmatrix} \\ \boldsymbol{A}_{2,1}(\boldsymbol{r}) = \frac{Gm_1}{||\boldsymbol{\rho}_1||^3}\left(\frac{3\boldsymbol{\rho}_1\boldsymbol{\rho}_1^T}{||\boldsymbol{\rho}_1||^2} - \boldsymbol{I}_3\right) + \frac{Gm_2}{||\boldsymbol{\rho}_2||^3}\left(\frac{3\boldsymbol{\rho}_2\boldsymbol{\rho}_2^T}{||\boldsymbol{\rho}_2||^2} - \boldsymbol{I}_3\right) - [\tilde{\boldsymbol{n}}][\tilde{\boldsymbol{n}}] \end{gathered} \tag{6}$$

Control is implemented as an impulsive $\delta\boldsymbol{v}$ which instantaneously changes spacecraft velocity. Because the chaser is the controlled vehicle, and the target is operating independent of the chaser, the control update is given by Equation 7. To assess the ideal performance of the developed guidance laws it is assumed there are no control error sources.

$$\boldsymbol{x}^{+} = \boldsymbol{x}^{-} + \boldsymbol{B}\delta\boldsymbol{v} \tag{7}$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{0}_3 & \boldsymbol{I}_3 & \boldsymbol{0}_3 & \boldsymbol{0}_3 \end{bmatrix}^T$$

B. Measurement Model

Bearing measurements are represented in many ways, but one common method is an azimuth and elevation angle pair as defined by Equation 8. Note, since the chaser is the controlled vehicle, it is assumed that it produces the measurements. Thus, the relative range vector is becomes $\boldsymbol{\rho} = \boldsymbol{r}_t - \boldsymbol{r}_c$, and its projections along the CR3BP axes are ρ_i, ρ_j, ρ_k.

$$\boldsymbol{y} = \begin{bmatrix} \tan^{-1}\left(\frac{\rho_j}{\rho_i}\right) & \sin^{-1}\left(\frac{\rho_k}{||\boldsymbol{\rho}||}\right) \end{bmatrix}^T \tag{8}$$

The measurement sensitivity matrix, defined as $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}} = \boldsymbol{H}$, is then given by Equation 9. Note, the partials with respect to the chaser are simply the negative of the partials with respect to the target due to the symmetric nature of relative measurements.

$$\begin{gathered} \boldsymbol{H} = \begin{bmatrix} -\boldsymbol{H}_{az} & \boldsymbol{0}_{1,3} & \boldsymbol{H}_{az} & \boldsymbol{0}_{1,3} \\ -\boldsymbol{H}_{el} & \boldsymbol{0}_{1,3} & \boldsymbol{H}_{el} & \boldsymbol{0}_{1,3} \end{bmatrix} \\ \boldsymbol{H}_{az} = \frac{1}{\rho_i^2 + \rho_j^2} \begin{bmatrix} -\rho_j & \rho_i & 0 \end{bmatrix} \\ \boldsymbol{H}_{el} = \frac{1}{\sqrt{||\boldsymbol{\rho}||^2 - \rho_k^2}} \begin{bmatrix} \frac{-\rho_i\rho_k}{||\boldsymbol{\rho}||^2} & \frac{-\rho_j\rho_k}{||\boldsymbol{\rho}||^2} & \frac{-\rho_k^2}{||\boldsymbol{\rho}||^2} + 1 \end{bmatrix} \end{gathered} \tag{9}$$

It is assumed that the measurements are subject to independent, identically, distributed Gaussian noise that is equivalent for both azimuth and elevation angles. The one-sigma angular uncertainty is set to 10 micro-rad to match previous optical navigation studies [4].

C. Problem Geometry and Variation

The objective of this work is to develop guidance laws to obtain information about the range between the chaser and target. Thus, having a geometric understanding of how maneuvers relate to optical measurement deviations to generate range information is critical. Figure 2 illustrates how a maneuver leads to change in angle at a future epoch. Given this geometry, Equation 10 relates range to its corresponding angle via the law of sines.

$$\rho_2 = \frac{\sin(\theta_1)}{\sin(\delta\theta)}\delta r \tag{10}$$

The first order variation of the law of sines can be taken to derive a relationship for a change in range given a change in parameters, as done in Equation 11. Then the square of the variation of range can be used to approximate uncertainty in a range estimate given the parameters.

$$\Delta\rho_2 = \frac{\sin(\theta_1)}{\sin(\delta\theta)}\Delta\delta r + \frac{\cos(\theta_1)}{\sin(\delta\theta)}\delta r \, \Delta\theta_1 - \frac{\sin(\theta_1)\cos(\delta\theta)}{\sin^2(\delta\theta)}\delta r \, \Delta\delta\theta \tag{11}$$

Taking the square of this variation is actually another method to derive the error of range estimates presented in previous bearing-only guidance literature [9]. To obtain the range error the prior work assumed the variation of position and perturbation angle are perfectly known, which translates to $\Delta\delta r$ and $\Delta\theta_1$ being zero in this notation. Then, note that $\delta\theta = \theta_2 - \theta_1$ and thus if the uncertainty in both angles are equivalent then its variation is $\Delta\delta\theta^2 = \Delta\theta^2$. Plugging this all in and squaring, an analog to the previously derived range error is obtained in Equation 12. One final note is that error given here includes an absolute value which ensures that the range uncertainty is positive definite as expected.

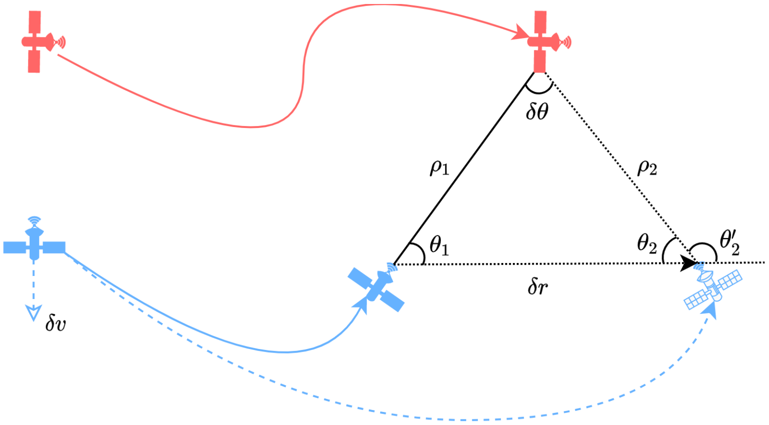

Fig. 2 Geometric change in angle due to a maneuver. The red craft is the target, and the blue is the controlled chaser. Solid lines are nominal trajectories, and the dashed line is the deviated trajectory.

$$\sigma_{\rho_2} \approx \frac{\rho_2}{|\tan(\delta\theta)|}\sigma_\theta \tag{12}$$

With the relevant geometry established, it is clear that maximizing the change in line-of-sight or decreasing range will reduce the range estimate uncertainty. This agrees with intuition as both of these effectively work to triangulate the target position given optical measurements. Moving forwards, Equation 11 will be used as the basis for range variational calculations, and Equation 12 will be used as the range estimate uncertainty.

III. Guidance Strategies

A. Heuristic Policy

From the geometric analysis it is understood that minimizing range uncertainty requires maximizing a change to the line-of-sight and minimizing range. This knowledge can be used to develop a useful heuristic policy given two primary assumptions: that the range between the craft is large and the coast time after a maneuver is short. From this, the small angle approximation for the change in angle can be appropriately incorporated. Also, the change in position can be linearly approximated as $\delta \boldsymbol{r} = \delta \boldsymbol{v}\ t$.

Given the aforementioned assumptions, the optimal policy to minimize range uncertainty from Equation 12 is to maneuver nearly perpendicular to the line-of-sight vector at the time of the maneuver. This can be verified numerically with ease by finding the optimal perturbation angle given a fixed ratio range ratio $\frac{\delta r}{\rho_1}$. The result of this optimization for the 2 dimensional simplified system is given in Figure 3. Thus, a candidate heuristic policy is to maneuver perpendicular to the line-of-sight such that $\theta_1 = \frac{\pi}{2}$.

If the heuristic policy is applied, and the original assumptions hold, then various simplification can be made to the range variation from Equation 11. First, the small angle approximation $\delta\theta \ll 1$ and perpendicular maneuver removes the $\Delta\theta_1$ term, as well as sine and cosine functions. Then linear motion is substituted and the equation is squared to approximate range uncertainty. Finally, the small angle and linear motion assumptions produce the relationship $\rho_1\delta\theta = \delta v\ t$, which can be used to substitute for $\delta\theta$. These operations result in Equation 13 and state that range uncertainty is a function of: range, maneuver magnitude, time, and measurement uncertainty.

$$\sigma_{\rho_\perp} \approx \frac{\rho_1^2}{\delta v\ t}\sqrt{\sigma_\theta^2 + \frac{t^2}{\rho_1^2}\sigma_{\delta v}^2} \tag{13}$$

The resultant range uncertainty equation states that range uncertainty is inversely proportional to maneuver magnitude and coast time. This conclusion agrees with intuition as larger maneuvers that can coast longer will create greater changes in geometry. Furthermore, maneuver errors are neglected for this paper to assess ideal guidance performance and therefore $\sigma_{\delta v}$ is zero. Thus, maneuvers perpendicular to the line-of-sight are adopted as a heuristic policy and the range information generated from such as a maneuver is approximated by Equation 14.

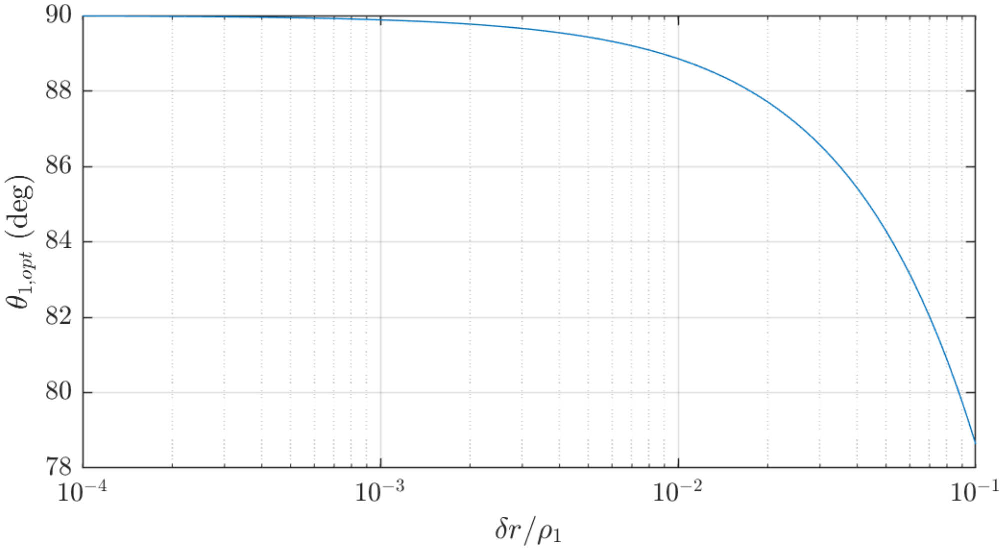

Fig. 3 Optimal perturbation angle to minimize Equation 12 for the 2 dimensional simplified system from Figure 2.

$$\sigma_{\rho_\perp} \approx \frac{\rho_1^2}{\delta v\ t}\sigma_\theta \tag{14}$$

The heuristic policy to maneuver perpendicular to the line of sight is useful for several reasons. First, it creates a baseline that is extremely easy to calculate and can be used to seed optimal solvers that require an initial guess. Second, the various assumptions made allowed for simplification of the problem so that range uncertainty was approximated as a function of maneuver magnitude. Finally, it provides a sanity check for optimal solvers since the optimal solution should approach the heuristic under the given assumptions.

B. Soft Actor-Critic

In addition to the heuristic strategy, we seek to train a model to find new maneuver strategies to reduce range uncertainty. Ideally, we'd like to find strategies that move away from the fixed Δv in the heuristic policy to find policies of equivalent or better range uncertainty reduction with less propellant expended. To this end, we implement Soft Actor-Critic (SAC), an off-policy actor-critic algorithm based on the maximum entropy RL framework [11, 12].

A number of algorithms were briefly studied and considered. Double Q Learning and other extensions to DQN have been shown to outperform humans in Atari games, but are limited to discrete state and action spaces [13]. Trust Region Policy Optimization and Proximal Policy Optimization allow for continuous domains, but implementing the advantage function was perceived to be challenging compared to the explicit actor-critic methods [14, 15]. SAC was chosen due to the resources available, ability to handle continuous domains, interest to the authors, and no perceived weaknesses.

In short, SAC implements deep neural networks to approximate a policy function (the actor, mapping the system state to a stochastic action) and a value function Q (the critic, mapping state and action space to an expected value). SAC also seeks to maximize entropy at each visited state, encouraging exploration by the agent while still maximizing rewards.

The actor policy, parameterized by ϕ, has a loss function specified by Equation 15 to enable gradient descent.

$$J_\pi(\phi) = \alpha log(\pi_\phi(\tilde{a}_t|s_t)) - Q_\theta(s_t, \tilde{a}_t) \tag{15}$$

$$\tilde{a}_t = tanh(\pi_\phi(\cdot|s_t)) \tag{16}$$

Where $\tilde{a}_t$ is a resampling from the policy that has be reparameterized to allow for ease of differentiation, and α is a temperature hyperparameter.

The critic value function, parameterized by θ, is approximated by two dueling neural networks Q_1 and Q_2, with target functions Q_{targ1} and Q_{targ2} to improve training stability. Training the critic involves minimizing the soft Bellman residual, meaning the loss function is given by Equation 17

$$J_Q(\theta) = \frac{1}{2}(Q_\theta(s_t, a_t) - (r(s_t, a_t) + \gamma(1 - d)(\min_{i=1,2} Q_{\theta targ,i}(s_{t+1}, \tilde{a}_{t+1}) - \alpha log(\pi_\phi(\tilde{a}_{t+1}|s_{t+1})))))^2 \tag{17}$$

$$\tilde{a}_{t+1} = \pi_\phi(\cdot|s_{t+1}) \tag{18}$$

Where r is the reward function, γ is the discount factor, and d is a binary factor set to 1 if the next state is terminal and 0 otherwise.

Our implementation of SAC relies on extending the POMDPs.jl package to define the formation flying problem [16]. The problem is defined as a Markov Decision Process (MDP), with the state fully observable, for simplicity. The state is defined as the position and velocity of both the target and chase spacecraft, the covariance of the chaser's estimate of the both spacecraft, and the time elapsed since the beginning of the scenario. The action space for the chase vehicle is a continuous Δv vector, which can be applied at the beginning of each one-hour time step.

The transition function is a deterministic propagation beginning with the application of the maneuver specified by the action, and ending after one hour. The dual spacecraft state is propagated according to Equations 5-(6). The covariance is propagated as a consider covariance, using the Kalman update equations with a measurement model specified in Equations 8 - 9. The number of hours elapsed is incremented by one.

The reward is specified by a positive reward given for reduction of the range uncertainty, given at the end of the scenario, and negative rewards given for the chaser's position deviation away from its free flight path (also given at the end of the scenario) and for total Δv expended.

The input vector for the state into the policy and value nets is a subset of the state: the chaser position and velocity, time elapsed, and the range uncertainty.

Training was performed in epochs of 100 episodes each. A circular buffer of 100,000 training data points was used, from which 1,000 randomly selected data points were used for training after each epoch. A learning rate of 0.0003 was used to match the implementation by Haarnoja, et al.

IV. Simulation and Results

A. Mission Design

To assess the various guidance policies a standardized simulation is created. In this simulation the target craft is placed on the 9:2 synodic Near Rectilinear Halo Orbit (NRHO) to mimic the Lunar Gateway, and the chaser is located on Quasi-Periodic Orbit (QPO) about the 9:2 NRHO. Leveraging the QPO ensures naturally bounded relative motion, which is ideal for an initial staging point before further proximity operations. In all cases the chaser must coast for a period before it is allowed to preform guidance maneuvers to improve its filter estimate for 12 hours, and then must return to its nominal trajectory after the designed guidance period.

The simulation starts at apoapsis because it is well known that maneuvers near periapsis can easily cause major trajectory dispersions, while maneuvers around apoapsis are well behaved. Thus, the simulation begins at apoapsis with a large initial uncertainty that is allowed to settle for 12 hours before the guidance maneuvers are performed. The 1-sigma initial uncertainty is 100 km and 1 m/s along each axis for both crafts. The guidance period follows the initial coast for 12 hours which provides sufficient time for the viewing geometry to change and for the filter to gain additional range information. Finally, at the end of the guidance period the chaser must return to its nominal and preforms a targeting maneuver to bring it back in another 24 hours. A note here is that the heuristic policies will burn at the start of the guidance period and simply coast till the end. This should be nearly optimal given a set amount of fuel to burn because it imitates bang-bang control.

For the initial guidance results a consider covariance analysis is executed, which provides several benefits. First, it removes random variables from the filter performance so that single runs can be directly compared to each other. Second, consider covariance allows for quick iteration of simulation design and fast data generation. Finally, it effectively sets a best case scenario for filter performance. In combination, consider covariance is well suited for comparing the different guidance laws because each simulation is fast, exactly repeatable, and produces an ideal performance.

B. Heuristic Policy

The heuristic policy is to maneuver perpendicular to the line of sight. This generates an admissible control set, and does not explicitly define a control. Therefore, an uncertainty quantification analysis of the admissible control is carried out to characterize the effectiveness of any allowable control from the heuristic. The results are compared to a baseline simulation with no maneuver, and a Monte-Carlo analysis of all possible control inputs which effectively identifies all possible results. For these simulations the chaser is allows 1 m/s of control which will be completely used at the start of the guidance period.

Figure 4 shows the evolution of the range uncertainty projection for the baseline, heuristic, and all possible control policies. The black dotted line is the natural range projection estimate if no maneuver is performed, with dots representing measurement updates. The solid red line is the average heuristic result, with the dashed red lines representing the upper and lower bounds. The solid blue line is the average Monte-Carlo of all possible control, with dashed lines representing the bounds again.

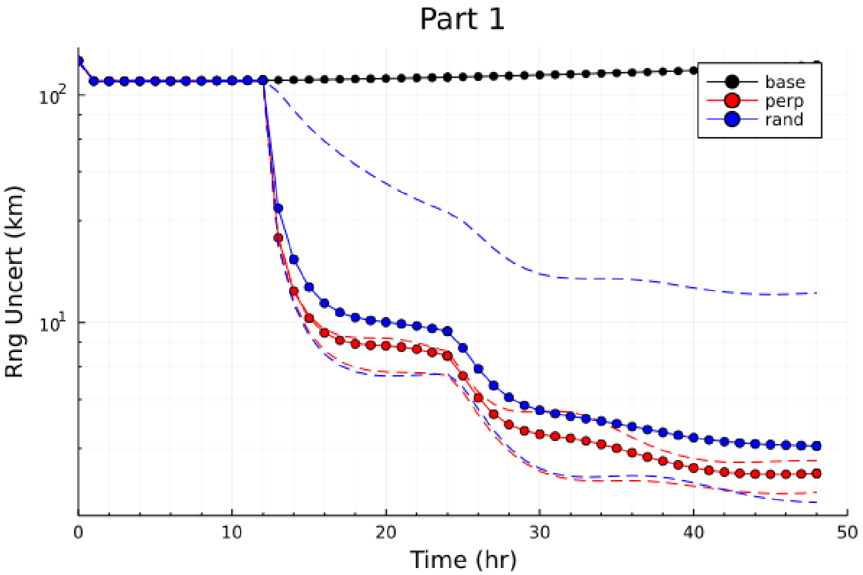

Fig. 4 Range uncertainty evolution for no maneuver in black, heuristic in red, and Monte-Carlo of all possible control in blue.

It is quite clear from these results that not only does the heuristic approximate the bottom bound, it also has a significantly improved upper bound. It is also evident that any maneuver performed in this scenario will improve range estimate uncertainty, but the heuristic guarantees to be within 4 km of the maximum range uncertainty reduction which improves the estimate by an order of magnitude. Finally, it is important to identify that while the first measurement has a significant impact on range uncertainty, all subsequent natural measurements produce little to no range information which is expected given optical measurements. Overall, the heuristic appears to be good policy to initialize control guesses on as desired.

C. Soft Actor-Critic

The policy found from Soft Actor-Critic began with training the actor and critic neural network. Training was performed in rounds of 1000 epochs, with 100 episodes each. At first, training was performed primarily on simulations running the heuristic policy, with an epsilon-greedy policy gradually taking over.

While the heuristic incorporates a total of three maneuvers, the trained policy is allowed to maneuver every hour beginning at the same 12 hour mark. The intent was to allow the training to discover new strategies, whether in new directions, magnitudes, or timing.

After 5000 epochs, the trained policy had nearly replicated the heuristic performance. The learning curve of the summed rewards achieved by the policy can be found on the left of Figure 5, showing a gradual but steady progression from random performance to roughly matching the heuristic. The range uncertainty over time achieved by the best

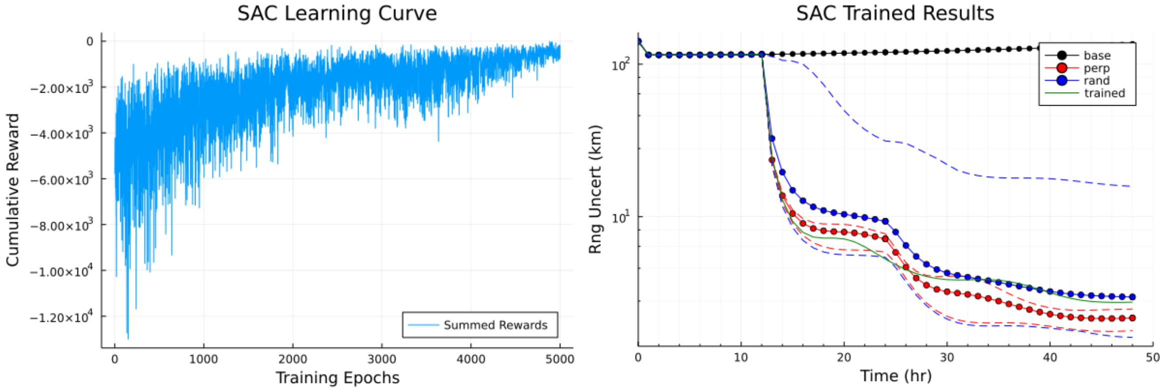

Fig. 5 The left plot is the learning curve for the Soft Actor-Critic policy over the first 5000 training epochs. The right plot is the range uncertainty evolution, comparing the heuristic policies in red and blue to the trained actor in green.

policy over the 5000 epochs is shown on the right of Figure 5. This trained policy consumed 4.41 m/s of Δv, compared to 3 m/s for the heuristic.

At this point, training diverged down two paths. The first path attempted to allow the policy to gain more experience with only minor tweaks to utilize an epsilon-greedy strategy more and to slightly increase the penalty for using propellant. The second attempted to change the reward structure to better incentivize the desired behavior by rewarding the actor for decreased range uncertainty on every time step instead of only at the terminal state and by limiting the maximum Δv. Both paths found results that were again in line with the heuristic bounds, without out-performing it as the authors intended.

V. Conclusion

Cislunar space is an exciting frontier of space development. Performing end-to-end autonomous navigation using only optical sensors is a light, low power, and cheap way to stay in formation around the moon. By adjusting coarse a satellite can change its line of sight and/or decrease its actual range which leads to a more accurate range estimate. A heuristic policy was developed that is optimal when the range between the two craft is large and maneuvers are small. This heuristic was then used to train a soft actor-critic algorithm which shows promising progress towards reaching the identified lower bound. We believe with further training and reward shaping that the policy would converge towards the lower bound.

Future work could continue along several avenues. First, more time can be dedicated to the soft actor-critic method as mentioned above. Second, a Monte-Carlo Tree Search with progressive widening could be implemented. Finally, the guidance methods developed here can be incorporated into a station keeping policy to further justify these maneuvers.

VI. Contributions and Release

- **The authors grant permission for this report to be posted publicly.**
- Jesse Greaves: Wrote system definition, dynamics, measurements, and updates. Created considered covariance analysis to simulate filter performance. Assessed heuristic policy.
- Andrew Palski: Investigated reinforcement learning algorithms. Wrote the MDP and Soft Actor-Critic implementations, trained the policy and value networks.
- Kevin Yevak: Investigated RL algorithms recommended. Worked on MCTS with Voronoi Progressive Widening as an alternative to Soft-Actor-Critic.

References

[1] *Artemis Plan: NASA's Lunar Exploration Program Overview*, NASA, September 2020. URL `https://www.nasa.gov/sites/default/files/atoms/files/artemis_plan-20200921.pdf`.

[2] Hill, K., and Born, G. H., "Autonomous interplanetary orbit determination using satellite-to-satellite tracking," *Journal of guidance, control, and dynamics*, Vol. 30, No. 3, 2007, pp. 679–686.

[3] Greaves, J. A., and Scheeres, D. J., "Relative Estimation in the Cislunar Regime using Optical Sensors," 2021.

[4] Bradley, N., Olikara, Z., Bhaskaran, S., and Young, B., "Cislunar Navigation Accuracy Using Optical Observations of Natural and Artificial Targets," *Journal of Spacecraft and Rockets*, Vol. 57, No. 4, 2020, pp. 777–792.

[5] Christian, J. A., and Lightsey, E. G., "Review of options for autonomous cislunar navigation," *Journal of Spacecraft and Rockets*, Vol. 46, No. 5, 2009, pp. 1023–1036.

[6] Grzymisch, J., and Fichter, W., "Observability criteria and unobservable maneuvers for in-orbit bearings-only navigation," *Journal of Guidance, Control, and Dynamics*, Vol. 37, No. 4, 2014, pp. 1250–1259.

[7] Woffinden, D. C., and Geller, D. K., "Observability criteria for angles-only navigation," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 45, No. 3, 2009, pp. 1194–1208.

[8] Grzymisch, J., and Fichter, W., "Optimal rendezvous guidance with enhanced bearings-only observability," *Journal of Guidance, Control, and Dynamics*, Vol. 38, No. 6, 2015, pp. 1131–1140.

[9] Woffinden, D. C., and Geller, D. K., "Optimal orbital rendezvous maneuvering for angles-only navigation," *Journal of guidance, control, and dynamics*, Vol. 32, No. 4, 2009, pp. 1382–1387.

[10] Ceresoli, M., Zanotti, G., and Lavagna, M., "Bearing-Only Navigation for Proximity Operations on Cis-Lunar Non-Keplerian Orbits," *72nd International Astronautical Congress (IAC 2021)*, 2021, pp. 1–10.

[11] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S., "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *CoRR*, Vol. abs/1801.01290, 2018. URL `http://arxiv.org/abs/1801.01290`.

[12] Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., and Levine, S., "Soft Actor-Critic Algorithms and Applications," Tech. rep., 2018.

[13] Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M. G., and Silver, D., "Rainbow: Combining Improvements in Deep Reinforcement Learning," *CoRR*, Vol. abs/1710.02298, 2017. URL `http://arxiv.org/abs/1710.02298`.

[14] Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P., "Trust Region Policy Optimization," *CoRR*, Vol. abs/1502.05477, 2015. URL `http://arxiv.org/abs/1502.05477`.

[15] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O., "Proximal Policy Optimization Algorithms," *CoRR*, Vol. abs/1707.06347, 2017. URL `http://arxiv.org/abs/1707.06347`.

[16] Egorov, M., Sunberg, Z. N., Balaban, E., Wheeler, T. A., Gupta, J. K., and Kochenderfer, M. J., "POMDPs.jl: A Framework for Sequential Decision Making under Uncertainty," *Journal of Machine Learning Research*, Vol. 18, No. 26, 2017, pp. 1–5. URL `http://jmlr.org/papers/v18/16-300.html`.