

Space Objects Maneuvering Detection and Prediction via Inverse Reinforcement Learning

Richard Linares*

University of Minnesota, Minneapolis, NM, 55455

Roberto Furfaro†

University of Arizona, Tucson, AZ, 85721

Abstract

This paper uses inverse Reinforcement Learning (RL) to determine the behavior of Space Objects (SOs) by estimating the reward function that an SO is using for control. The approach discussed in this work can be used to analyze maneuvering of SOs from observational data. The inverse RL problem is solved using the Feature Matching approach. This approach determines the optimal reward function that a SO is using while maneuvering by assuming that the observed trajectories are optimal with respect to the SO's own reward function. This paper uses estimated orbital element data to determine the behavior of SOs in a data-driven fashion. Simple proof-of-concepts results are shown for a simulation example.

1 Introduction

Space Situational Awareness (SSA) has many definitions depending on the goal at hand, but in general it involves collecting and maintaining knowledge of all space objects (SOs) orbiting the Earth and the space environment. This task is becoming more difficult as the number of objects currently tracked by the U.S. increases due to breakup events and improving tracking capabilities [1]. The Space Surveillance Network (SSN) is tasked with maintaining information on over 22,000 objects, 1,100 of which are active, with a collection of optical and radar sensors. Determining physically significant characteristics, i.e. attributes, that go beyond simple orbital states is a key objective which is required for protecting space capabilities and achieving SSA. For example, the SSN catalog currently includes radar cross-section and a non-conservative force parameter, analogous to a ballistic coefficient, which provides additional SO characterization information beyond position and velocity. Future SSA systems will have to be capable of building a much more detailed picture of SO attributes in order to maintain better knowledge of their characteristics, which ultimately may lead to better tracking capabilities. This work develops a method for characterizing the behavior of SOs from observational data.

Traditional measurement sources for SO tracking, such as radar and optical, have been shown to provide information on SO characteristics. These measurements have been shown to be sensitive to shape [2, 3], attitude [2, 4, 5], angular velocity [6], and surface parameters [7, 8]. State-of-the-art in the literature has been advanced over the past decades and in recent years seen the development of multiple model [2, 9], nonlinear state estimation [4–6], and full Bayesian inversion [10] approaches for SO characterization. This work provides a fundamentally different and novel way of studying the SO characterization problem via maneuver estimation. Methods for detecting maneuvers of Space Objects (SOs) can provide useful indications and warning information for Space Situational Awareness. In particular, SOs may maneuver unexpectedly or fail to perform station keeping maneuvers [11]. These types of anomalous maneuvering of SOs can indicate a threat to other active SOs. Therefore, it is important to predict and understand maneuvering patterns.

The SSN data are provided mainly from ground based radar and optical systems and although these systems meet almost all tactical needs for cooperative earth satellites, there are serious shortcomings for

*Assistant Professor, Department of Aerospace Engineering and Mechanics. Email: rlinares@umn.edu, Member AIAA.

†Associate Professor, Department of Systems & Industrial Engineering. Email: robertof@email.arizona.edu

HEO and GEO cooperative earth satellite and all noncooperative earth orbiting satellite [12]. The ground based sensor network is limited in that there exist large geographical gaps in the SNN coverage and sensor detection sensitivity is low for HEO and GEO satellites and other distant SOs. This results in a slow response to intentional and unintentional noncooperative maneuvers and break up events, reducing SSA capabilities and the ability of gathering data for earth satellites in higher orbits. This work seeks to develop methods for noncooperative maneuver prediction and modeling.

This work discusses the use of inverse Reinforcement Learning (RL) to learn the behavior of Space Objects (SOs) from observed orbital motion. The behavior of SOs is estimated using inverse RL to determine the reward function that each SO is using to control. Since SOs with have the capability of maneuvering are controlled to achieve a particular goal which is mission driven, maneuvering can be very subjective and only a data-driven learning approach can reveal the true goal. It is also important to determine what type of behavior a SO is using and if this behavior changes. Inverse RL approaches use optimal control principles to learn what reward function is being used by an agent given observations. The simplest inverse RL approach, discussed in Ref. [14], solves for the reward function using a weighted sum of features. The weights determined from the inverse RL calculation are the representation for the reward function the expert is using.

The estimated reward function weights can be used to determine the type of behavior mode the SO is following and to classify the model based on libraries of behavior models. These weight vectors can be added to the state of SOs as a way to represent the policy that the SO is currently following and allow for the change of this policy over time and the behavior changes. Using the inverse RL approach we can formulate the optimal control problem as a Markov Decision Process (MDP) where we are not explicitly given a reward function. Rather, we are given observations of expert demonstrations for a given task and the goal is to estimate the reward function that the expert used to derive the demonstration trajectories. It is common to assume that the expert's actions are optimal with respect to the reward function the expert is using and our proposed approach makes this assumption. This work will investigate the Feature Matching Approach (FMA) [14] for solving for the expert's reward function. This paper studies the problem of finding potential features for using orbital element data. General features are developed using radial basis functions of the orbital element state.

The organization of this paper is as follows. First, the concept of reinforcement learning and Q-learning is introduced. Following this the inverse reinforcement learning approach is discussed and this is followed by an outline of a method for maneuver estimation. Finally, simulation results are provided for the proposed methods.

2 Reinforcement Learning

This section provides a brief introduction to reinforcement learning. Given a discrete-time system model, we can denote the state of the system at time step k by \mathbf{x}_k . The system dynamics provide the transition from \mathbf{x}_k to \mathbf{x}_{k+1} given \mathbf{u}_k , where $\mathbf{u}_k \in \mathcal{R}^\ell$ denotes the current control action, and this transition may be stochastic. Therefore, it is meaningful to represent this transition with a probability distribution $\mathbf{x}_{k+1} \sim p(\mathbf{x}_{k+1} | \mathbf{x}_k, \mathbf{u}_k)$ and $\mathbf{x}_k, \mathbf{x}_{k+1} \in \mathcal{R}^n$ denotes the current and next state, respectively. The actions are modeled probabilistically and are generated by a policy $\mathbf{u}_k \sim \pi(\mathbf{u}_k | \mathbf{x}_k)$ where the randomness in the policy can enable exploration of the policy space while also providing optimality for certain classes of control problems.

An agent (the maneuvering SO) has a current state $\mathbf{x}_k \in \mathcal{S}$ (orbital elements) at each discrete time step k and chooses an action \mathbf{u}_k according to a policy π . For the policy π , a reward signal r_k is given for a transition to a new state \mathbf{x}_{k+1} . The general objective of RL is to maximize an expectation over the discounted return, $J(\theta)$, given as:

$$J(\theta) = \mathbb{E}_{\pi_\theta} \{ r_k + \gamma r_{k+1} + \gamma^2 r_{k+2} + \dots \} \quad (1)$$

where $\gamma \in [0, 1)$ is a discount factor and θ are policy parameters. Q-learning is a popular RL method which defines a Q-function that represents the total reward or the total "cost-to-go" for a policy π [13]. Once the Q-function is determined, the action with the highest value or estimated total reward is taken at each time step. Therefore, the optimal policy can be determined using the optimal Q-function. The Q-function of a

policy π is:

$$Q^\pi(\mathbf{x}_k, \mathbf{u}_k) = \mathbb{E}_{\pi_\theta} \left\{ \sum_{i=k}^{\infty} \gamma^{i-k} r_i \right\} \quad (2)$$

Where the function estimates the total discounted reward for policy π from state \mathbf{x}_k assuming that action \mathbf{u}_k is taken and then all following actions are sampled from policy π . Q-network based methods use neural networks parameterized by θ to represent $Q^\pi(\mathbf{x}_k, \mathbf{u}_k; \theta)$, but we drop the dependency notation for simplicity [13]. Q-networks are optimized by minimizing the following loss function:

$$\mathcal{L}(\theta) = \left(r_k + \gamma \max_{\mathbf{u}_{k+1}} Q^\pi(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - Q^\pi(\mathbf{x}_k, \mathbf{u}_k) \right)^2 \quad (3)$$

This equation uses the Bellman optimality condition [13] to relate $Q^\pi(\mathbf{x}_k, \mathbf{u}_k)$ to $Q^\pi(\mathbf{x}_{k+1}, \mathbf{u}_{k+1})$, and this equation can be optimized using stochastic gradient descent. Given a reward function, RL can be used to determine a Q-function which is optimal with respect to this reward function.

3 Inverse Reinforcement Learning

The basic principle behind inverse RL is to find the expert's reward function, $r_k(\mathbf{x}_k, \mathbf{u}_k)$, that explains the expert's behavior given the observations. Reference [14] introduced a feature based reward function where the reward function is a linear combination of nonlinear features. If the set of features is sufficiently rich, this assumption is fairly unrestrictive. For example, in the extreme case, where a separate feature is used for each state-action pair, fully general reward functions can be approximated. Using this model, the reward function is written as,

$$r_k(\mathbf{x}, \mathbf{u}_k) = \mathbf{w}^T \phi(\mathbf{x}_k, \mathbf{u}_k) \quad (4)$$

where $\phi(\mathbf{x}_k, \mathbf{u}_k) = [\phi_1(\mathbf{x}_k, \mathbf{u}_k), \dots, \phi_n(\mathbf{x}_k, \mathbf{u}_k)]^T$ are user defined functions and the reward function can be estimated by finding the vector \mathbf{w} . The feature Matching Approach (FMA) [14] solves for the vector $\mathbf{w} = [w_1, \dots, w_n]^T$ by matching the expectation of the features given the expert policy, π , and a policy based on the estimation reward, π^* . The feature expectation is then written as,

$$\mu(\pi) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k \phi(\mathbf{x}_k, \mathbf{u}_k) \right\} \quad (5)$$

The FMA finds the vector \mathbf{w} by computing $\mu(\pi)$ for a set of policies $\pi^{(i)}$ and ensuring that expert is always optimal over all policies $\pi^{(i)}$. In other words, the expert is assumed to be optimal with respect to its own reward function and therefore all other policies must be sub-optimal. The FMA searches the space of $\pi^{(i)}$ in a way to guaranty rapid convergence to solution for \mathbf{w} . This principle can be expressed compactly in the following condition,

$$E_{\pi^*} \left\{ \sum_{k=0}^{\infty} \gamma^k \mathbf{w}^T \phi(\mathbf{x}_k, \mathbf{u}_k) \right\} \geq E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k \mathbf{w}^T \phi(\mathbf{x}_k, \mathbf{u}_k) \right\} \quad (6)$$

This condition ensures that the expert is optimal with respect to the estimated reward function. Then the weights can be found by solving the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \|\mathbf{w}\|_2^2 \\ \text{subject to} \quad & \mathbf{w}^T \boldsymbol{\mu}(\pi^*) \geq \mathbf{w}^T \boldsymbol{\mu}(\pi^{(i)}) + 1 \end{aligned}$$

This optimization problem is formulated from the $\mathbf{w}^T \boldsymbol{\mu}(\pi^*) \geq \mathbf{w}^T \boldsymbol{\mu}(\pi^{(i)})$ condition while removing the trivial solution $\mathbf{w} = \mathbf{0}$ by adding the minimization of the squared L_2 norm of the vector.

$$\begin{aligned} \min_{\mathbf{w}, \tau} \quad & \tau \\ \text{subject to} \quad & \mathbf{w}^T \boldsymbol{\mu}(\pi^*) \geq \mathbf{w}^T \boldsymbol{\mu}(\pi^{(i)}) + \tau, \quad \|\mathbf{w}\|_2^2 \leq 1 \end{aligned}$$

4 Maneuver Detection

This section discusses the maneuver detection and orbit determination approached used for this work. We developed an orbit determination approach for maneuver detection based on the Square-root Unscented Kalman Filter (S-UKF) [15] which was chosen for its robustness and ability to handle nonlinear functions. The orbit determination method developed for this work has the option of using the Simplified General Perturbations model (SGP4) or a custom high fidelity orbital propagator. The S-UKF estimation approach also has the option of using process noise for accounting for unmodeled disturbance forces. The S-UKF approach developed for this work can work in two modes, in a filtering model or in a batch smoothing model. The orbit determination process requires the following inputs; 1) RA and Dec measurement sequences, 2) Time of observation in Julian date format, 3) Earth orientation parameter file for observation times, 4) Location of ground site in Lat-Long and altitude (using wgs-84). While the orbit determination process provides the following outputs; 1) Mean motion (rad/min), 2) eccentricity, 3) inclination (Deg), 4) Argument of perigee (Deg), 5) RA of the ascending node (Deg), 6) mean anomaly (Deg).

The proof-of-concept results of the S-UKF approach are shown in Figure 2. There are two types of maneuvers, maneuvers occurring during observation segments and maneuvers occurring between observation segments. The second type of maneuver is much more likely since persistent observations are not readily available. The detection of maneuvers during observation segments can be thought as occurring “while we are watching,” otherwise the effect of the maneuver has to be inferred from orbital dynamics. This work focuses on detection of maneuvers that occur between observation segments that can be thought of as occurring “when we aren’t watching.”

The proof-of-concept approach uses an initial orbit determination process to look for “outliers” in the data and these “outliers” are flagged as maneuvers. We assume that enough observations are made available to determine the orbit of the SO. The proposed approach uses the residuals of the orbit determination process to determine if a maneuver has occurred. The measurement residuals are given by the following:

$$\tilde{\mathbf{y}}_k = \mathbf{y}_k - \hat{\mathbf{y}}_k \quad (7)$$

where $\tilde{\mathbf{y}}_k$, \mathbf{y}_k , and $\hat{\mathbf{y}}_k$ are the measurement residual, measured values, and the estimated predicted measurement, respectively. The estimated predicted measurement is provided by a Square-root Unscented Kalman Filter (S-UKF) orbit determination approach. The S-UKF approach also provides a covariance for the state which can be used to calculate the predicted measurement residual covariance or the innovation, S_k . Then under normal orbit determination conditions (where all assumptions are met) the scaled norm of the residual is Chi-Squared distributed and is given by

$$\epsilon_k = \tilde{\mathbf{y}}_k^T S_k^{-1} \tilde{\mathbf{y}}_k \quad (8)$$

Therefore, maneuvers can be detected using the ϵ_k variable and testing whether this variable is Chi-Squared distributed. Under the no maneuvering cases all of the assumptions of the S-UKF are met and therefore the variable ϵ_k is close to being Chi-Squared distributed but when a maneuver occurs the dynamic model in the S-UKF will not provide a good estimate since it does not incorporate maneuvers and therefore, when maneuver occur the variable ϵ_k is not Chi-Squared distributed. The statistics of this test are improved by considering a set of residuals over a window of time. This work uses two methods for selecting the set of residuals, a sliding window method and a fading-memory moving window. The sliding window ϵ_k^s calculation is given by

$$\epsilon_k^s = \sum_{j=k-s+1}^k \epsilon_j \quad (9)$$

The fading-memory moving window ϵ_k^ρ calculation is given by

$$\epsilon_k^\rho = \sum_{j=1}^k \rho^{k-j} \epsilon_j = \rho \epsilon_{k-1}^\rho + \epsilon_k \quad (10)$$

Both variables, ϵ_k^ρ and ϵ_k^s , are used in this work for calculating measurement residuals for the S-UKF and comparing these statistics to a Chi-Squared distribution. If the measurement residuals do not match the

Chi-Squared distribution then a maneuver is detected. The two maneuver detection approaches used for this work have been shown to be robust to actual outliers and measurement errors through simulation proof-of-concept studies shown in the next section.

5 Simulation Results

This section discusses the initial proof-of-concept results for the proposed maneuver modeling approach. Two cases are considered in this section. The first case discussed the initial results for the maneuver detection approach. This approach uses a S-UKF to detect maneuvers from observations. After maneuvers are detected the inverse reinforcement learning approach is used to learn the reward function the produced the maneuver. The learned reward function can then be used to estimate the behavior of SOs.

5.1 Maneuver Detection Results

Simulation studies were conducted to test the proposed maneuver detection approach. The proof-of-concept simulations have shown that the proposed approach can detect maneuvers using optical satellite measurements. Figure 1 shows the initial proof-of-concept results achieved. The Chi-Squared maneuver detection approach was tested using simulated observations from a near-GEO satellite. A numerical orbital simulation was developed which included 2-body forces, solar radiation pressure, and J2 perturbations. A ground based sensor was used and right ascension and declination measurements of a GEO satellite were simulated.

Two simulation cases are considered, a East-West GEO thrusting and a North South GEO thrusting. Two maneuvers simulated East-West (E-W) (2 cm/s Figure 1) and North-South (N-S) (150 cm/s Figure 1). It is well known that N-S is more detectable, whereas E-W is smaller and may take more time to be detected from observation data. The N-S example looked at for this work has a relatively larger maneuver and therefore should be detectable. The fading memory and fixed window moving sums are used for detecting the N-S maneuver using a 99.5% confidence level. The results for the fading memory and fixed window approaches are shown in Figure 1. Both approaches used a Chi-Squared test 99.95% confidence level. More statistics are developed by considering the proposed moving window and fading memory Chi-Squared variables. The results in Figure 1 show that the fading memory approach can detect the maneuver quicker than the fixed window approach, although both work well.

5.2 Learning Reward Function from Maneuvers

The simulation scenario considered for learning the reward function used a GEO stationary SO which is performing station keeping maneuvers to maintain a near GEO orbit. The reward model used for generating the simulated observations is shown in Figure 3(a). The SO was given reward for maintain a near GEO orbit around it's nominal location and additional reward for "hitting" two spots at -2.5 Deg south and -1 Deg west of it's nominal GEO location and additionally 2.5 Deg south and 1 Deg west of it's nominal GEO location. These location can be thought of as communication windows that the satellite would like to "hit" during its orbit. Given this true reward function, Q-learning is used to generate simulated trajectories. Then these simulated trajectories are provided as observations for the inverse reinforcement learning approach. For these initial proof-of-concept results perfect measurements of the satellites orbital elements are assumed. Given the assumed orbital elements, the actions for the SO will be the orbital element differences for a given maneuver. To avoid excessive maneuvering a negative reward of -10 is given for each maneuver. Then the FMA approach discussed in the earlier section is used for estimating the reward function. The features used for the reward functions learning are a collection of radial basis function in the North-South (N-S) and East-West (E-W) space. The basis functions are given by

$$\phi_i(\mathbf{x}_k) = \exp \left\{ -\frac{1}{2\lambda} (\mathbf{r}(\mathbf{x}_k) - \boldsymbol{\mu}_i)^T (\mathbf{r}(\mathbf{x}_k) - \boldsymbol{\mu}_i) \right\} \quad (11)$$

where $\mathbf{r}(\mathbf{x}_k)$ is the radial N-S and E-W distance for the desired GEO SO location and the parameters $\boldsymbol{\mu}_i$ are the means of the radial basis functions. The inverse reinforcement learning method then learns the reward function as a linear combination of the radial basis functions given in Eq. (11). The estimated reward function is then given by the learned weights w_i for the i^{th} basis. A total of 25 basis functions are used

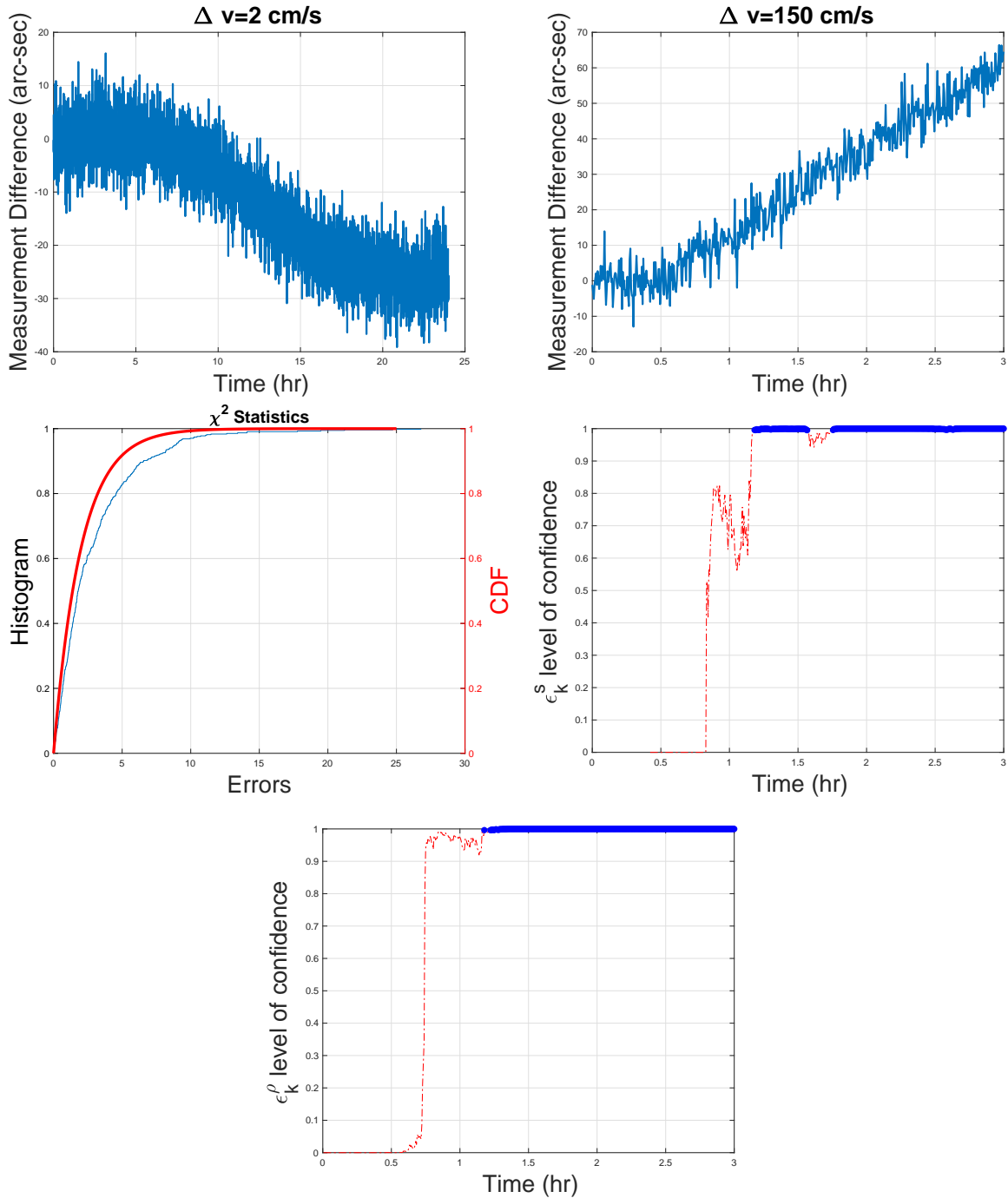


Figure 1: Observations for East-West Burn (upper left), Observation for North-South Burn (upper right), Non-Maneuvering Chi-Square Test (middle left), ϵ_s Based Confidence Level (middle right), ϵ_p Based Confidence Level (lower left).

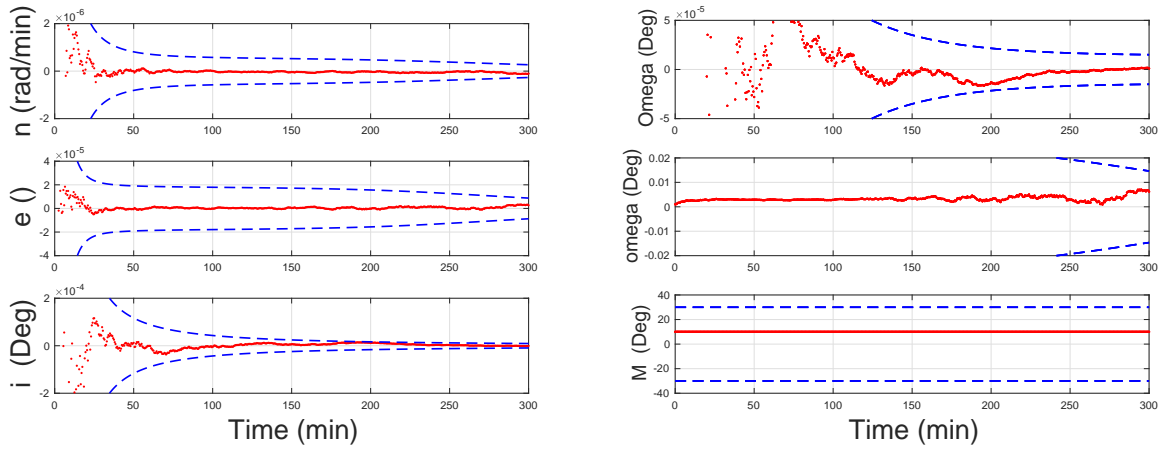


Figure 2: S-UKF Estimation Results Using SGP-4 Propagation Model.

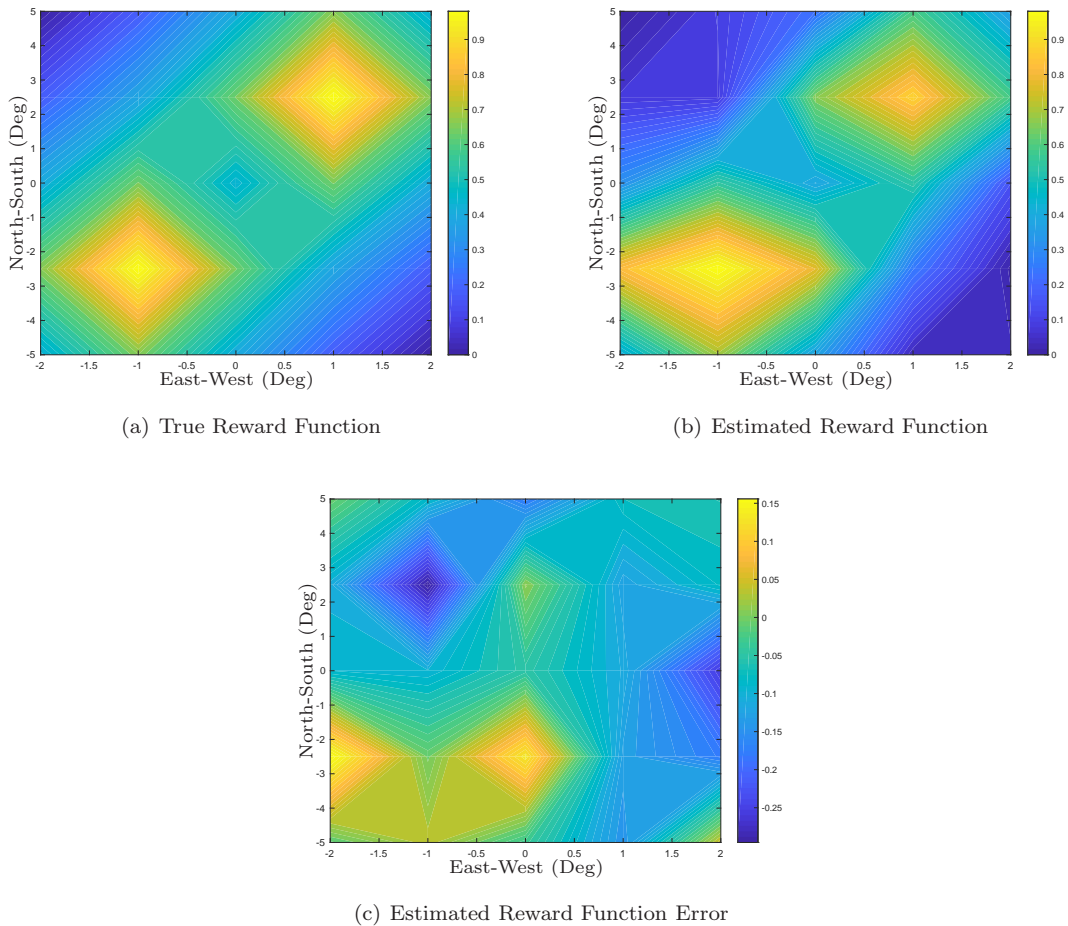


Figure 3: Estimated vs. True Reward Function for GEO Maneuvering SO

to approximate the reward function. The means of the basis functions are positioned on a 5×5 grid over 2 Degr in E-W and 5 Degr in N-S. The λ parameters are selected to be the same for all basis functions where $\lambda = 1$ was used for this work. Figure 3(b) shows the reward function that was estimated directly from the

observed trajectories using the FMA approach. The estimated reward function captured the general trend of the maneuvering behavior. From the figure it can be seen that although the FMA method estimates the lower-left peak it misses the upper-right peak of the true reward function. This can be seen directly from Figure 3(c) which shows the error between the estimated and the true reward function. From Figure 3(c) it can be seen that the maximum error in the reward function is 0.15. Additionally, future work will consider more advanced inverse RL methods for estimating the reward function of an SO from observations.

6 Conclusion

This paper considered the problem of determining the behavior of a SO from observational data using inverse reinforcement learning. Two aspects of the problem were considered, a maneuver detection approach and a reward model determination approach. A method for detecting maneuvers from observational data was discussed that uses the Square root Unscented Kalman Filter (S-UKF) to “smoothen” the state trajectories. This method then determines the time of maneuvers using a Chi-Squared test on the orbit determination residuals. If the orbit determination residual are above a given value in terms of the z-score a maneuver is indicated. Finally, with this maneuvering indication the state vectors can be used to estimate the Δv required for such a maneuver. Given the Δv and the state at which these Δv occurred, the inverse reinforcement learning approach can be used to estimate the reward function that a given SO is using. This work considered a simulated case of an SO in GEO which is maneuvering to maintain a given GEO stationary box. The true reward function was specified and simulation data was generated for a hypothetical SO which is following this true reward function. Finally, the reward function was estimated using a linear combination of features. The nonlinear basis functions used for this work was a set of radial basis functions. The estimated reward function approximated the true reward function well and good performance was shown for the proposed approach.

References

- [1] House, W., “National Space Policy of the United States of America,” *Retrieved from https://www.whitehouse.gov/sites/default/files/national_space_policy_6-28-10.pdf*, 2010.
- [2] Linares, R., Jah, M. K., Crassidis, J. L., and Nebelecky, C. K., “Space Object Shape Characterization and Tracking Using Light Curve and Angles Data,” *Journal of Guidance, Control, and Dynamics*, Vol. 37, No. 1, 2013, pp. 13–25.
- [3] Hall, D., Calef, B., Knox, K., Bolden, M., and Kervin, P., “Separating Attitude and Shape Effects for Non-resolved Objects,” *The 2007 AMOS Technical Conference Proceedings*, 2007, pp. 464–475.
- [4] Jah, M. and Madler, R., “Satellite Characterization: Angles and Light Curve Data Fusion for Spacecraft State and Parameter Estimation,” *Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conference*, Vol. 49, Wailea, Maui, HI, Sept. 2007, Paper E49.
- [5] Holzinger, M. J., Alfriend, K. T., Wetterer, C. J., Luu, K. K., Sabol, C., and Hamada, K., “Photometric attitude estimation for agile space objects with shape uncertainty,” *Journal of Guidance, Control, and Dynamics*, Vol. 37, No. 3, 2014, pp. 921–932.
- [6] Linares, R., Shoemaker, M., Walker, A., Mehta, P. M., Palmer, D. M., Thompson, D. C., Koller, J., and Crassidis, J. L., “Photometric Data from Non-Resolved Objects for Space Object Characterization and Improved Atmospheric Modeling,” *Advanced Maui Optical and Space Surveillance Technologies Conference*, Vol. 1, 2013, p. 32.
- [7] Linares, R., Jah, M. K., Crassidis, J. L., Leve, F. A., and Kelec, T., “Astrometric and photometric data fusion for inactive space object feature estimation,” *Proceedings of 62nd International Astronautical Congress, International Astronautical Federation*, Vol. 3, 2011, pp. 2289–2305.
- [8] Gaylor, D. and Anderson, J., “Use of Hierarchical Mixtures of Experts to Detect Resident Space Object Attitude,” *Advanced Maui Optical and Space Surveillance Technologies Conference*, Vol. 1, 2014, p. 70.

- [9] Wetterer, C. J., Linares, R., Crassidis, J. L., Kececy, T. M., Ziebart, M. K., Jah, M. K., and Cefola, P. J., “Refining space object radiation pressure modeling with bidirectional reflectance distribution functions,” *Journal of Guidance, Control, and Dynamics*, Vol. 37, No. 1, 2013, pp. 185–196.
- [10] Linares, R. and Crassidis, J. L., “Resident Space Object Shape Inversion via Adaptive Hamiltonian Markov Chain Monte Carlo,” *AAS/AIAA Space Flight Mechanics Meeting*, No. AAS Paper 2016-514, Napa, CA, Feb 2016.
- [11] Kececy, T., Hall, D., Hamada, K., and Stocker, D., “Satellite maneuver detection using Two-line Element (TLE) data,” *Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conference*, Maui Economic Development Board (MEDB) Maui, HA, 2007.
- [12] Naka, F., Canavan, G., Clinton, R., Judd, O., and Pensa, A., “Space Surveillance, Asteroids and Comets, and Space Debris. Volume 1: Space Surveillance,” Tech. rep., SCIENTIFIC ADVISORY BOARD (AIR FORCE) WASHINGTON DC, 1997.
- [13] Sutton, R. S. and Barto, A. G., *Reinforcement learning: An introduction*, Vol. 1, MIT press Cambridge, 1998.
- [14] Abbeel, P. and Ng, A. Y., “Apprenticeship learning via inverse reinforcement learning,” *Proceedings of the twenty-first international conference on Machine learning*, ACM, 2004, p. 1.
- [15] Van Der Merwe, R. and Wan, E. A., “The square-root unscented Kalman filter for state and parameter estimation,” *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on*, Vol. 6, IEEE, 2001, pp. 3461–3464.