AAS 25-479

INFORMATION-DRIVEN CISLUNAR SPACE OBJECT REACQUISITION

Disip Chaturvedi*, G. Andrew Siciliano*, and Keith A. LeGrand†

The surge in activities within the cislunar domain has underscored the need for advanced algorithms to detect and track noncooperative space objects. Unknown orbital maneuvers performed during sensor coverage gaps can lead to significant trajectory deviations, making the redetection of such objects challenging, especially in the presence of false detections. This paper proposes sensor tasking algorithms for the reacquisition of a previously detected cislunar space object that has deviated from its nominal trajectory. The problem is modeled as a partially observable Markov Decision Process (POMDP) with a random finite set (RFS)-theoretic observation model. An information-driven approach based on Monte Carlo tree search (MCTS) with double progressive widening (DPW) is considered for online POMDP planning. The performance of the proposed approach is compared against that of random and greedy search strategies in terms of time taken to reacquire the lost object. Through Monte Carlo trials, it is demonstrated that the MCTS-based approach consistently outperforms these strategies.

1 Introduction

The problem of sensor tasking entails designing and developing methods for searching or gaining custody of space objects (SOs) and tracking or maintaining custody of known SOs. As interest in cislunar space operations grows, there is an increasing need for advanced, efficient algorithms to detect and track cislunar space objects (CSOs). The problem is challenging due to the vast search space and complex dynamics of the cislunar environment. Moreover, factors such as sensing limitations, sensor noise, non-cooperative CSO, and visibility conditions exacerbate the difficulty. Ground-based observers are limited by restricted observation windows, occultation, and atmospheric conditions, while space-based sensors [1] face challenges due to limited observational range and the kinematic constraints of the spacecraft carrying them.

These issues have been addressed in the sensor management literature, with approaches typically categorized as (i) Top-Down/Control-theoretic methods, and (ii) Bottom-Up/Heuristic-based methods [2]. Top-down approaches formulate the problem as an optimization task aimed at minimizing the overall uncertainty in the observed SO states. In contrast, the Bottom-up approaches divide sensor management into a series of structured and sequential sub-tasks with their sequence governed by combining heuristics (rule-based) and mission-oriented optimization techniques. Recent developments in sensor tasking [3–7] utilize information-theoretic objective functions, offering a systematic framework for quantifying and optimizing the uncertainty.

In space situational awareness (SSA), the traditional sensor management methods include optical surveys and follow-up to catalog space objects [8–10]. More recently, heuristic-based and optimization-based methods for cataloging and maintenance have been proposed. Heuristic-based target search strategies such as set-path search [11, 12] and meta-heuristics methods [4, 13, 14], have gained prominence in sensor tasking problems due to their computational efficiency and analytical clarity. Top-down methods for cataloging and maintaining custody encompass a variety of approaches including but not limited to: greedy optimization [7, 15], genetic algorithms [16], Monte Carlo tree search (MCTS) [17, 18], Deep Reinforcement Learning (DRL) [19, 20], and integer programming [21]. These techniques have also been adapted for sensor

*Ph.D. Student, School of Aeronautics & Astronautics, Purdue University, 701 W. Stadium Ave. West Lafayette, IN.
†Assistant Professor, School of Aeronautics & Astronautics, Purdue University, 701 W. Stadium Ave. West Lafayette, IN.

tasking in the cislunar regime [21–24]. Problems on searching for SOs have also been studied in [25–28]. In [18], dynamically feasible regions are sampled and MCTS is used to efficiently explore these regions for the time-optimal recovery of SOs. In [29], known maneuver bounds are used to determine a reachable set for a maneuvering SO lost in space and a probability of detection greedy policy is used to search these sets.

Despite significant advances in search and tracking, the problem of searching for SOs remains challenging due to large continuous search space compounded by the presence of noise, missed detections, false alarms, and object appearance/disappearance. In this context, Monte Carlo methods provide an effective approach for probabilistically exploring the search space, offering a systematic way to account for uncertainties and enabling the development of search strategies that outperform myopic, suboptimal solutions. Drawing from the strengths of Monte Carlo methods and leveraging the inherent partially observable Markov Decision Process (POMDP) nature of sensor tasking problem, the solutions in [18, 22] employ MCTS with double progressive widening (DPW) for effective sensor motion planning in uncertain and continuous spaces. On the other hand, such state trajectory sampling approaches, preclude the usage of belief state-dependent rewards, including information gain functionals. Alternatively, belief-space approaches such as the particle filter tree with double progressive widening (PFT-DPW) [30] enable belief state-dependent rewards, yet have been limited to low-dimensional problems due to their reliance on particle filter based belief representations.

This research addresses a cislunar search-while-tracking problem using belief Markov decision process (BMDP) planning, generalizing the approach to incorporate random finite set (RFS) observation processes. This extension permits consideration of false alarms and missed detections in possible sensing outcomes. RFS theory [2] serves as a powerful tool, extending the Bayesian filtering framework to facilitate the detection and tracking of an unknown number of objects. The proposed approach is illustrated through a specific scenario involving the reacquisition of a lost cislunar object using a space-based sensor. In this case, the object's location uncertainty significantly exceeds the sensor's bounded field-of-view, necessitating multiple sensor reorientations to rediscover the object. A crucial constraint considered in this scenario is the limited slew capability of the sensor between time steps.

Section 2 formulates the CSO reacquisition problem as a stochastic optimal control problem utilizing RFS theory. The proposed algorithm for solving this optimal control problem is presented in Section 3. The tracking algorithm used is then described in Section 4. Section 5 underscores the algorithm's effectiveness in CSO reacquisition and evaluates its performance. Section 6 provides a summary of the research.

2 Problem Formulation

In this section, the details of the CSO motion are provided, and the formulation of reacquiring a lost cislunar space object as a stochastic optimal control problem is presented.

2.1 Spacecraft Dynamics

Let $\boldsymbol{x}_k \in \mathbb{X} \subseteq \mathbb{R}^6$ represent the object's state at time instant t_k which can be defined as

$$\boldsymbol{x}_k = [\boldsymbol{r}_k^\top \;\; \boldsymbol{v}_k^\top]^\top \tag{1}$$

where $\boldsymbol{r}_k = \begin{bmatrix} x & y & z \end{bmatrix}^\top$ is the position of the CSO with respect to the Earth-Moon barycenter ($\boldsymbol{o}_{\mathrm{EM}}$), and $\boldsymbol{v}_k$ is the velocity of the CSO in the synodic frame $\mathcal{O} : (\boldsymbol{o}_{\mathrm{EM}}, \hat{\boldsymbol{o}}_1, \hat{\boldsymbol{o}}_2, \hat{\boldsymbol{o}}_3)$. The circular restricted three-body problem (CR3BP) assumptions are adopted for simplicity. A CSO is considered whose equations of motion (EOMs) admit the Jacobi integral of motion. The non-dimensionalized EOM are given as

$$\ddot{x} - 2\dot{y} = \left.\frac{\partial U}{\partial x}\right|_{\boldsymbol{x}}, \qquad \ddot{y} + 2\dot{x} = \left.\frac{\partial U}{\partial y}\right|_{\boldsymbol{x}}, \qquad \ddot{z} = \left.\frac{\partial U}{\partial z}\right|_{\boldsymbol{x}} \tag{2}$$

where $U(x, y, z)$ is the non-dimensional pseudo-potential function $U(x, y, z)$ given as

$$U(x, y, z) = \frac{1-\mu}{r_{/\mathrm{Earth}}} + \frac{\mu}{r_{/\mathrm{Moon}}} + \frac{x^2 + y^2 + z^2}{2} \tag{3}$$

and μ is the non-dimensional ratio of the mass of the Moon to the sum of the masses of the Moon and the Earth. The quantities $r_{/\text{Earth}}$ and $r_{/\text{Moon}}$ are the distances of the CSO from the Earth and the Moon, respectively. The non-dimensional Jacobi integral is defined as

$$C(\boldsymbol{x}) = 2U(x, y, z) - \boldsymbol{v}^\top \boldsymbol{v} \tag{4}$$

For estimation and sensor tasking, the continuous-time dynamics are approximated as a discrete-time nonlinear system

$$\boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}) + \boldsymbol{w}_{k-1} \tag{5}$$

where $\boldsymbol{f}(\boldsymbol{x})$ is the CR3BP deterministic solution flow and $\boldsymbol{w}_{k-1}$ is a zero-mean Gaussian white process noise with covariance Q_{k-1}.

2.2 Measurement Model

The CSO is searched for using a space-based observer with an optical sensor rigidly mounted onto it. At instant t_k, the space-based observer has an associated body frame $\mathcal{B}_k : (\boldsymbol{r}_{o,k},\ \hat{\boldsymbol{b}}_{1,k},\ \hat{\boldsymbol{b}}_{2,k},\ \hat{\boldsymbol{b}}_{3,k})$ where $\boldsymbol{r}_{o,k}$ is the position vector of the observer with respect to the Earth-Moon barycenter, and $\{\hat{\boldsymbol{b}}_{1,k} \quad \hat{\boldsymbol{b}}_{2,k} \quad \hat{\boldsymbol{b}}_{3,k}\}$ are orthonormal basis vectors of the frame $\mathcal{B}_k$. The sensor is assumed to be mounted along the $\hat{\boldsymbol{b}}_{1,k}$ direction. The remaining basis frame vectors of $\mathcal{B}_k$ are defined as

$$\hat{\boldsymbol{b}}_{2,k} = -\frac{\hat{\boldsymbol{b}}_{1,k} \times \hat{\boldsymbol{o}}_3}{|\hat{\boldsymbol{b}}_{1,k} \times \hat{\boldsymbol{o}}_3|} \qquad\qquad \hat{\boldsymbol{b}}_{3,k} = \hat{\boldsymbol{b}}_{1,k} \times \hat{\boldsymbol{b}}_{2,k} \tag{6}$$

The observer spacecraft state $\boldsymbol{\xi}_k$ at t_k consists of the observer's position $\boldsymbol{r}_{o,k}$ and attitude parametrized as a vector-first quaternion $\boldsymbol{q}_{\mathcal{B}/\mathcal{O},k}$. The observer's orbital trajectory is known and governed by the CR3BP EOMs. The observer's attitude is assumed to be fully controllable. The observer's state determines the sensor boresight direction and its rectangular pyramid field-of-view (FoV) $\mathcal{S}_k(\boldsymbol{\xi}_k)$. This dependence of FoV on the observer state $\boldsymbol{\xi}_k$ is implicitly assumed throughout the paper and is suppressed for brevity henceforth. The CSO detection depends only on the position states $\boldsymbol{r}_k \in \mathbb{X}_s$, hence the FoV is modeled as a bounded and compact subset in $\mathbb{X}_s$. The decomposition $\mathbb{X}_s$ is such that $\mathbb{X}_s \times \mathbb{X}_v = \mathbb{X}$ recovers the full state space.

The presence of the CSO in a FoV $\mathcal{S}_k$ is expressed by the generalized indicator function

$$1_{\mathcal{S}_k(\boldsymbol{\xi}_k)}(\boldsymbol{x}_k) = \begin{cases} 1, & \text{if } \boldsymbol{r}_k \in \mathcal{S}_k(\boldsymbol{\xi}_k) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The object detection is assumed to be random and characterized by the probability function

$$p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) = 1_{\mathcal{S}_k(\boldsymbol{\xi}_k)}(\boldsymbol{x}_k) \cdot p_{D,k}(\boldsymbol{r}_k; \boldsymbol{r}_{o,k}) \tag{8}$$

where $p_{D,k}(\boldsymbol{r}_k; \boldsymbol{r}_{o,k})$ is the probability of object detection at the target position for an unbounded FoV. Let $\boldsymbol{r}_{t/o,k}$ be the relative position vector of the CSO with respect to the observer given as

$$\boldsymbol{r}_{t/o,k} = \frac{\boldsymbol{r}_k - \boldsymbol{r}_{o,k}}{\|\boldsymbol{r}_k - \boldsymbol{r}_{o,k}\|} \tag{9}$$

When an object is detected, the sensor generates a noisy angular measurement $\boldsymbol{z}_k \in \mathbb{Z} \subseteq \mathbb{R}^2$ according to the likelihood function $p_k(\boldsymbol{z}_k|\boldsymbol{x}_k)$ conditioned on the existence of an object with state $\boldsymbol{x}_k$ and generation of an observation $\boldsymbol{z}_k$. This angular measurement is produced per the nonlinear measurement model

$$\boldsymbol{z}_k = [\theta_k,\ \phi_k]^\top = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{\nu}_k = \left[\tan^{-1}\left(\frac{\boldsymbol{r}_{t/o,k}\cdot\hat{\boldsymbol{b}}_{1,k}}{\boldsymbol{r}_{t/o,k}\cdot\hat{\boldsymbol{b}}_{2,k}}\right) \quad \sin^{-1}\left(\boldsymbol{r}_{t/o,k}\cdot\hat{\boldsymbol{b}}_{3,k}\right)\right]^\top + \boldsymbol{\nu}_k \tag{10}$$

where θ_k and ϕ_k are the azimuth and elevation of the CSO at time t_k, measured with respect to the sensor's local horizon, and $\boldsymbol{\nu}_k$ is a zero-mean Gaussian white noise with covariance R_{k-1}. The CSO is detected

only if it is within the FoV, well-illuminated, not occluded by the Earth/Moon, and is not located in the Moon exclusion angle. The observations are modeled as a RFS to account for random spurious and missed detections. The observation RFS at t_k is given as

$$Z_k = Z(t_k) = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m(k)}\} \tag{11}$$

where $m(k) = |Z_k|$ is the unknown time-varying measurement cardinality. If the intensity of false alarms is sufficiently small, the conditional observation process is approximately Bernoulli with state-conditioned likelihood

$$f_k(Z|\boldsymbol{x}_k; \boldsymbol{\xi}_k) = \begin{cases} 1 - p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k), & Z = \emptyset \\ p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) p_k(\boldsymbol{z}_k|\boldsymbol{x}_k), & Z = \{\boldsymbol{z}_k\} \end{cases} \tag{12}$$

The observer's attitude is governed by the information-driven controller developed in this paper. The achievable sensor orientations within a single time step are limited by kinematic and operational constraints of the observer spacecraft. These constraints are incorporated into the sensor motion planning by considering a maximum slew rate, which defines the admissible control set $\mathcal{U}_k$ for the observer's attitude at each decision epoch. Because the observer's orbital trajectory is fixed, the FoV $\mathcal{S}_k$ and the attitude $\boldsymbol{q}_{\mathcal{B}/\mathcal{O}}$ are equivalent representations up to a boresight rotation and thus are used interchangeably. The sensor's discrete-time attitude dynamics are given as

$$\boldsymbol{q}_{\mathcal{B}/\mathcal{O},k} = \boldsymbol{f}_s(\boldsymbol{q}_{\mathcal{B}/\mathcal{O},k-1}, \boldsymbol{u}_k), \qquad \boldsymbol{u}_k \in \mathcal{U}_k \tag{13}$$

where $\boldsymbol{u}_k$ is the control action. In this work, the relative attitude itself is used as the control action.

2.3 Space-Based Sensor Search-while-Tracking

The space-based sensor search-while-tracking problem constitutes a POMDP. It is assumed that a previously detected CSO was lost in space and the only available information at the start of the search time t_0 is the probability density function (pdf) $p_0(\boldsymbol{x})$ of the CSO state. Let the posterior CSO state pdf after the instant t_{k-1} be $p_{k-1}(\boldsymbol{x}_{k-1}|Z_{1:k-1})$, where $Z_{1:k-1}$ denotes the time series of measurement RFSs from time t_1 to t_{k-1}. The corresponding sensor FoVs are $\mathcal{S}_{1:k-1}$. The distribution p can be replaced with multi-object pdfs while considering multi-target cases. The problem entails determining sensor the control sequence $\boldsymbol{u}_{k:k+N-1}$ or equivalently the FoV sequence $\mathcal{S}_{k:k+N-1}$ over a finite horizon N that leads to CSO detection and tracking. A finite planning horizon is considered to mitigate suboptimal solutions from myopic greedy approaches.

To systematically reduce uncertainty in the CSO state, an information-driven tasking problem is formulated in which sensor actions expected to yield the highest information content are sought. The information gained from a given observation is measured using the Kullback-Leibler divergence (KLD), which quantifies the similarity between the posterior distribution and prior distribution. The KLD, also known as the *relative entropy*, between two integrable densities p and q is given by

$$D_{\mathrm{KL}}(p\|q) = \int p(\boldsymbol{x}) \cdot \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \tag{14}$$

where the support of $p(\cdot)$ is assumed to be contained within the support of $q(\cdot)$, and the information theoretic convention of $0 \log \frac{0}{0}$ is adopted [31]. When $q(\cdot)$ and $p(\cdot)$ represent the prior and posterior densities, the KLD is a measure of *information gain*. The expected information gain over possible measurements Z is also known as the *mutual information*, given by

$$I(\boldsymbol{x}; Z) = \mathrm{E}_Z[D_{\mathrm{KL}}(p(\boldsymbol{x}|Z) \,\|\, p(\boldsymbol{x}))] \tag{15}$$

where for convenience the same notation is used for both random variables and variates. The expectation in (15) is with respect to the RFS Z and thus

$$\mathrm{E}_Z[D_{\mathrm{KL}}(p(\boldsymbol{x}|Z) \,\|\, p(\boldsymbol{x}))] = \int D_{\mathrm{KL}}(p(\boldsymbol{x}|Z) \,\|\, p(\boldsymbol{x})) f(Z) \delta Z \tag{16}$$

where the set differential notation δZ implies that the integral in (16) is a set integral, which is defined for an arbitrary set function $f(Y)$ as

$$\int f(Y)\delta Y \triangleq \sum_{n=0}^{\infty} \frac{1}{n!} \int f(\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}) d\boldsymbol{y}_1 \ldots d\boldsymbol{y}_n \tag{17}$$

Thus, the sensor tasking problem formulated as a stochastic optimal control problem aimed at maximizing the expected information gain over the planning horizon is given as

$$\begin{aligned} \max_{\boldsymbol{u}_{k-1:k+N-1}} & \mathrm{E}_{Z_{k:k+N}} \left[\sum_{i=k}^{k+N} D_{\mathrm{KL}} \left(p_i(\boldsymbol{x}|Z_{1:i}) \,\|\, p_{i|i-1}(\boldsymbol{x}|Z_{1:i-1}) \right) \right] \\ \text{subject to } & \boldsymbol{u}_i \in \mathcal{U}_i; \;\; \forall\, i \in \{k-1, \ldots, k+N-1\} \\ & \boldsymbol{q}_{\mathcal{B}/\mathcal{O},i} = \boldsymbol{f}_s(\boldsymbol{q}_{\mathcal{B}/\mathcal{O},i-1}, \boldsymbol{u}_i) \end{aligned} \tag{18}$$

The KLD implicitly depends on the control $\boldsymbol{u}_i$ through the dependence of the distribution $p_i(\cdot)$ on the measurement RFS Z_i, which in turn is dependent on the sensor attitude $\boldsymbol{q}_{\mathcal{B}/\mathcal{O},i}$.

3 Methodology

The stochastic optimization problem (18) presents several challenges. The highly nonlinear and chaotic nature of three-body dynamics in cislunar space (3), coupled with the nonlinearity of the measurement model (10), results in non-Gaussian distributions of the CSO state. Consequently, a closed-form formulation of the reward function D_{KL} is unavailable. An effective solution approach to non-Gaussian nonlinear filtering is Gaussian mixture (GM) filtering [32], where pdfs are approximated as a weighted sum of L Gaussian mixands:

$$p(\boldsymbol{x}) \approx \sum_{\ell=1}^{L} w^{(\ell)} \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}^{(\ell)}, \boldsymbol{P}^{(\ell)}) \tag{19}$$

where $w^{(\ell)}$, $\boldsymbol{m}^{(\ell)}$, and $\boldsymbol{P}^{(\ell)}$ denote the weights, means, and covariances respectively of the ℓ^{th} component, and $\mathcal{N}(\boldsymbol{x}; \boldsymbol{m}, \boldsymbol{P})$ represents a Gaussian distribution with mean $\boldsymbol{m}$ and covariance $\boldsymbol{P}$. The GM representations of the non-Gaussian prior and posterior distributions facilitate accurate computation of KLD (D_{KL}) using Monte-Carlo integration. From the pdf $p_{k-1}(\boldsymbol{x}_{k-1}|Z_{1:k-1})$ at time t_{k-1}, the prior $p_{k|k-1}(\boldsymbol{x}_k|Z_{1:k-1})$ can be predicted and updated using the measurement to find the posterior $p_k(\boldsymbol{x}_k|Z_{1:k})$. The evolution of these pdfs follows an iterative process

$$\cdots \rightarrow p_{k-1}(\boldsymbol{x}_{k-1}|Z_{1:k-1}) \rightarrow p_{k|k-1}(\boldsymbol{x}_k|Z_{1:k-1}) \rightarrow p_k(\boldsymbol{x}_k|Z_{1:k}) \rightarrow \cdots \tag{20}$$

This paper employs two different GM based filters. The first is a high-accuracy adaptive filter that accounts for possible target maneuvers and is used as the primary tracking algorithm, as described in Section 4. The second filter is a lightweight GM filter used in the solution of the stochastic optimal control problem and is described in Section 3.2.

Another challenge arises from the attitude-based control input, which introduces a continuous and high-dimensional action space in the optimization. The high-dimensional characteristic gives rise to a combinatorial explosion in the number of possible actions as the dimensionality of the action space grows. Moreover, the continuous nature of the action space exacerbates this issue, effectively rendering the set of possible actions infinite within each dimension. These factors increase the complexities of optimization.

Another issue arises in the computation of the reward function in (18), which is the expected information gain over the planning horizon. The reward function is essentially a set integral over the RFS Z as seen in (16), which is generally intractable due to the infinite summation of single object integrals (17). This intractability necessitates an approximation for tractable computation of the reward function. Moreover, this

set expectation must be evaluated for every action sequence considered in the optimization. One approach to address this challenge is to consider predicted measurements from the prior and use Bayesian updates to construct potential posterior distributions. However, this approach introduces its own set of difficulties. The single-object measurement space $\mathbb{Z}$ is continuous, resulting in an infinite number of possible observations for a given action. This necessitates the development of efficient methods to explore the action and observation spaces in a computationally feasible manner.

MCTS emerges as powerful solution to these challenges. This heuristic search algorithm offers a computationally feasible approach for exhausting the vast state and action spaces. The strength of MCTS lies in its selective expansion strategy. Rather than attempting to exhaustively search through all possible actions — an intractable task in high-dimensional, continuous spaces — MCTS focuses computational resources on the most promising regions of the action space. MCTS can be adapted to continuous spaces through DPW that allows gradual expansion of the search tree preventing an explosion of the tree width in the initial steps. The key principle of DPW [33] is that as certain nodes prove more valuable, the algorithm explores more nodes in the continuous neighborhood around those nodes, essentially *widening* the tree for more promising nodes.

This paper addresses information-driven planning by leveraging belief-dependent rewards such as KLD, through the use of BMDP. A BMDP allows reformulation of a POMDP as a Markov decision process (MDP) where the state space is the space of all possible belief distributions over the underlying system states. Rather than directly traversing the infinite-dimensional belief space to identify valuable beliefs, the method employed here focuses on searching the action and observation spaces that lead to valuable beliefs. The action space represents all actions the observer spacecraft can take, while the observation space encompasses all potential observations that can be received as a result of those actions. To efficiently navigate these spaces, progressive widening is applied to actions and observations. This enables a more tractable exploration of the high-dimensional and continuous spaces inherent in the problem, while maintaining the ability to identify information-rich regions in the planning space.

3.1 Belief Space Planning

The key details of belief space planning using MCTS with DPW are now discussed. The high-level algorithm is a slightly generalized version of the PFT-DPW algorithm [30], refered to as the Bayes-DPW algorithm and describe in Algorithm 2. The key differences of the Bayes-DPW algorithm with respect to [30] include a GM-based belief state to accommodate higher dimensional state spaces and the generalization for RFS observations.

Algorithm 1 describes the procedure to construct the tree and find the optimal action. Algorithm 2 describes the SIMULATE procedure, which constitutes the main MCTS simulation and involves selection, expansion, rollout, and backpropagation, each of which is described in detail below.

Algorithm 1 MCTS Run

procedure RUN(n_0, N_{iter}, d_{max})
  for $i \in 1 : N_{\text{iter}}$ **do**
    SIMULATE(n_0, d_{max})
  end for
  return $\arg\max_{\boldsymbol{u}} Q(p, \boldsymbol{u})$
end procedure

The BMDP beliefs (p) are posterior pdfs represented as GMs as defined in (19). Given a posterior pdf (p), MCTS solves the BMDP by exploring potential actions $(\boldsymbol{u})$ that produce different belief trajectories over the planning horizon. The cumulative information gained along each of these trajectories informs the value $(Q(p, \boldsymbol{u}))$ of each explored action enabling the selection of the optimal action. The BMDP framework requires a *generative model* to stochastically generate these belief trajectories which is discussed in Section 3.2. The tree structure naturally accommodates this sequential decision-making process where depth from the root n_0 corresponding to belief p captures the temporal evolution of beliefs and breadth at each depth accounts

for exploration of different possible action-observation pairs. The search tree is constructed from the posterior pdf iteratively through N_{iter} simulations, where each simulation explores action-observation sequences up to a maximum depth $d_{\max}$ corresponding the planning horizon. Figure 1 shows the tree structure along with the key steps of the algorithm.

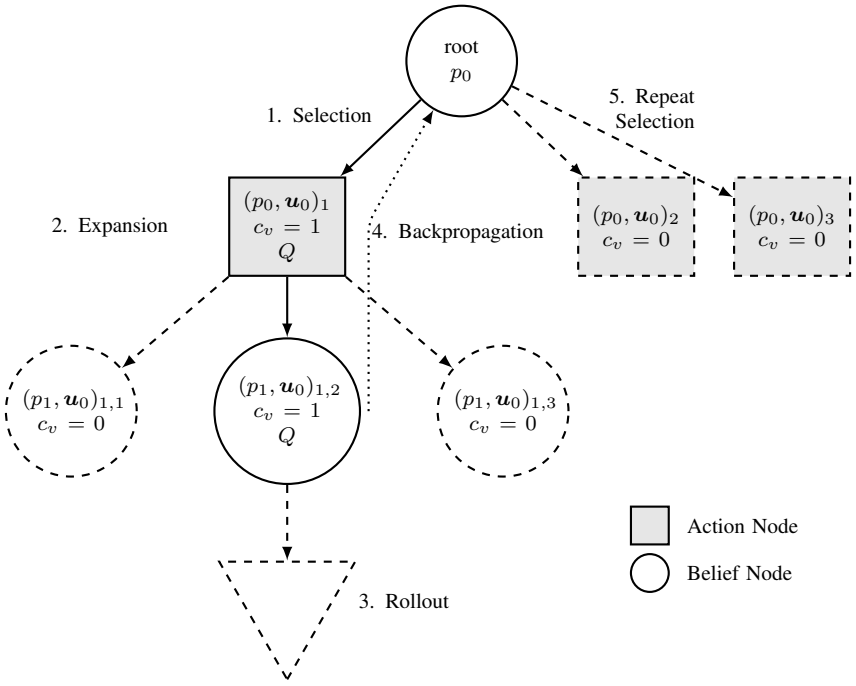


Figure 1: Monte Carlo tree search tree.

A **node** (n) forms the basic unit of the tree. The search tree alternates between action and observation nodes, indicating that the planning is carried out in the action and observation spaces. Each node maintains a count c_v of simulations that have passed through that node during tree construction. The *observation nodes* (n_o) correspond to a measurement RFS (Z) simulated from a sample of the prior belief and a selected action. The observation nodes maintain an updated belief based on the simulated measurement. The *root node* (n_0) is an observation node corresponding to the last true measurement and the corresponding posterior CSO state distribution. An *action node* (n_a) corresponds to a control action (u) that rotates the sensor pointing direction and determines the sensor FoV used to generate the measurement sample. Each action node maintains an empirical average of rewards accumulated from its resulting trajectories.

Selection in MCTS involves choosing an action for expanding the tree from the current observation node. The potential action nodes following an observation node are determined by the admissible control set based on the maximum permissible slew from the parent orientation state. The admissible control set of possible sensor orientations can be modeled as a continuous space or as a discrete set of orientations that produce non-overlapping FoVs. To explore such large action spaces systematically and in a computationally feasible manner, action progressive widening is used. The tree is allowed to gradually widen introducing new potential actions for the observation nodes with higher visit counts. This approach ensures comprehensive exploration over multiple iterations without exhaustively sampling the entire space at once. Although action progressive widening effectively handles both continuous and large discrete spaces, this paper adopts the discretized action space for simplicity. The selection of action from the possible actions for a given observation node is governed by a *tree policy* that balances the explore-exploit trade-off by balancing visits between promising nodes (exploitation) and less-explored nodes (exploration). The upper-confidence bound for trees (UCT) [33] is utilized as tree policy in the ACTIONSELECTION procedure, given its theoretical convergence properties and extensive validation in MCTS applications. From the admissible control set corresponding to a given

Algorithm 2 Bayes-DPW (generalized from [30])

procedure SIMULATE(n_o, d)
  if $d = 0$ **then**
    return 0
  end if
  $n_a \leftarrow$ ACTIONSELECTION(n_o) ▷ Select node n_a with action $\boldsymbol{u}$.
  if $|C(n_a)| \leq k_o c_v(n_a)^{\alpha_o}$ **then** ▷ $C(n_a)$ is the list of children of action node n_a.
    $p',\ Z,\ r,\ l(Z) \leftarrow$ GENERATIVEMODEL(p, n_a) ▷ p is the belief of parent of node n_a.
    $n_o \leftarrow$ OBSERVATIONNODE($p', Z, l(Z)$)
    $C(n_a) \leftarrow C(n_a) \cup \{n_o\}$
    DOROLLOUT $\leftarrow 1$ ▷ Rollout is required.
  else
    $n_0 \leftarrow$ likelihood based sample from $C(n_a)$
    DOROLLOUT $\leftarrow 0$ ▷ Rollout is not performed.
  end if
  if DOROLLOUT **then**
    $r_{\text{total}} \leftarrow r + \gamma \cdot$ ROLLOUT($n_o, d-1$)
  else
    $r_{\text{total}} \leftarrow r + \gamma \cdot$ SIMULATE($n_o, d-1$)
  end if
  $c_v(n_o) \leftarrow c_v(n_o) + 1$
  $c_v(n_a) \leftarrow c_v(n_a) + 1$
  $Q(p, \boldsymbol{u}) \leftarrow Q(p, \boldsymbol{u}) + \frac{r_{\text{total}} - Q(p, \boldsymbol{u})}{c_v(n_a)}$
  return r_{total}
end procedure

observation node (n_o), the child action node (n_a) whose action results in maximum UCT score is selected as [33]

$$n_a = \underset{n_a \in \text{ children of } n_o}{\arg\max} \; Q(p, \boldsymbol{u}) + \eta \sqrt{\frac{\log(c_v(n_o))}{c_v(n_a)}} \Bigg\} \tag{21}$$

where p is the belief associated with n_o and η is a hyper-parameter that balances exploration and exploitation. Lower values of η bias exploration towards actions yielding immediate rewards, while larger values induce excessive randomness in the search process. For a chosen N_{iter}, the value of η must be carefully tuned to ensure sufficient visitation across the action space.

A bias in action exploration arises from the order in which action nodes are added to the admissible set or populated in the tree, especially when multiple action nodes have the same UCT score. To avoid such biases, the actions are populated or selected greedily based on the expected number of measurements in the FoV corresponding to the action. When the greedy policy ties, a uniform selection is made.

The **expansion** step involves generating an observation for a selected action and adding the corresponding action-observation nodes to the tree for simulation. The observation is randomly generated according to the RFS generative model, as discussed in Section 3.2. Since the measurements are stochastically generated from a continuous distribution, the number of possible observations for a given belief and action pair is practically infinite. To facilitate a systematic sampling of the measurement space and prevent explosive tree growth, this work generalizes a process known as observation progressive widening for finite set observations. A new observation is generated for an action node only when the number of its existing children (representing previously simulated observations) is less than the widening threshold $k_o(c_v(n_a))^{\alpha_o}$, where $c_v(n_a)$ is the action node's visit count. The hyperparameter k_o determines the minimum number of observations to consider before widening is performed, and α_o governs how quickly the tree is allowed to widen. This allows focusing on sufficient initial exploration for different action nodes but gradually widening only the nodes that appear

promising in the long term. When a new observation is generated from an action node, the tree is expanded with a new observation node, initialized using OBSERVATIONNODE function, containing the corresponding updated belief for rollout. Otherwise, an existing observation is selected for simulation at the next depth through likelihood-weighted sampling, where the observation likelihoods $l(Z)$ are computed as detailed in Section 3.3.

When an existing action-observation pair is sampled, the **simulation** proceeds to the next depth. This recursive procedure expands the tree evaluating different trajectories by generating empirical reward samples at each depth. When a new action-observation pair is added to the tree, a **rollout** is performed to determine its long-term value. A rollout is a forward simulation up to the maximum depth without expanding the tree. The action selection policy used during a rollout, called the *rollout policy*, is typically simpler to allow for fast reward evaluation. In this work, two different rollout policies are considered. A *greedy rollout* selects the next action to immediately maximize the expected number of target-originated measurements in the FoV corresponding to the action. On the other hand, a *random rollout* chooses the next action by uniformly sampling from the list of possible actions. The function ROLLOUT runs the specified rollout policy.

Backpropagation updates the node statistics (visit counts and reward estimates) along the traversed path in the tree using the reward samples from rollout or simulation. For each node in the path, the visit count is incremented and the empirical average reward is updated with the new sample.

After tree construction, the action yielding the maximum value estimate at the root node is selected, as shown in Figure 1. The sensor pointing direction is then reoriented using the corresponding quaternion. The resulting measurement is collected and used to update the belief distribution CSO state. This process is continued until the target is confidently detected.

3.2 Generative Model

Having described the MCTS-based BMDP planning approach, the details of generative model required by the BMDP framework are presented in this section. For a given action and prior belief, the generative model stochastically generates an observation, updates the prior belief with this generated observation, and computes the corresponding reward. This procedure is represented by GENERATIVEMODEL function in Algorithm 2. Given the parent node posterior belief, its associated time-updated prior and the considered action $\boldsymbol{u}$, the observation is stochastically generated according to (12). The observation cardinality is first determined according to a random Bernoulli trial. The observation is missed with probability $1 - p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) d\mathbf{x}_k$. Otherwise, a singleton observation is sampled according to the unnormalized density $p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) p_k(\boldsymbol{z}_k | \boldsymbol{x}_k)$.

Belief Update The belief update incorporates measurement information into the prior CSO state distribution through Bayesian inference. Specifically, an adaptive Gaussian Mixture Bayes' filter [34] is employed to perform this update, accounting for state-dependent probability of detection. The posterior GM at t_k for the generated RFS measurement Z_k is given as

$$\begin{aligned} p_k(\boldsymbol{x}_k | Z_{1:k}) \propto \delta_{\emptyset}(Z_k) & \sum_{\ell=1}^{L_{k|k-1}} (1 - p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k)) w_{k|k-1}^{(\ell)} \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{m}_{k|k-1}^{(\ell)}, P_{k|k-1}^{(\ell)}) \\ & + (1 - \delta_{\emptyset}(Z_k)) \sum_{\ell=1}^{L_{k|k-1}} p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) w_{k|k-1}^{(\ell)} p_k(\boldsymbol{z}_k | \boldsymbol{x}_k) \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{m}_{k|k-1}^{(\ell)}, P_{k|k-1}^{(\ell)}) \end{aligned} \tag{22}$$

where $\delta_{\emptyset}(Z_k)$ is a set-generalized Kronecker delta function, $\boldsymbol{m}_{k|k-1}^{\ell}$, $P_{k|k-1}^{\ell}$ are parameters of the prior distribution expressed as a GM with $L_{k|k-1}$ components. The updated belief $p_k(\mathbf{x}_k | Z_{1:k})$ is approximated as a GM whose components are given in [34].

Negative Information When searching for a lost CSO, many observations will contain no object-originated detections. The absence of detections is in fact valuable information and is known colloquially as *negative information*. Negative information plays a critical role in CSO reacquisition and must be incorporated into the information state to avoid revisiting low probability regions [6, 7]. Negative information is incorporated

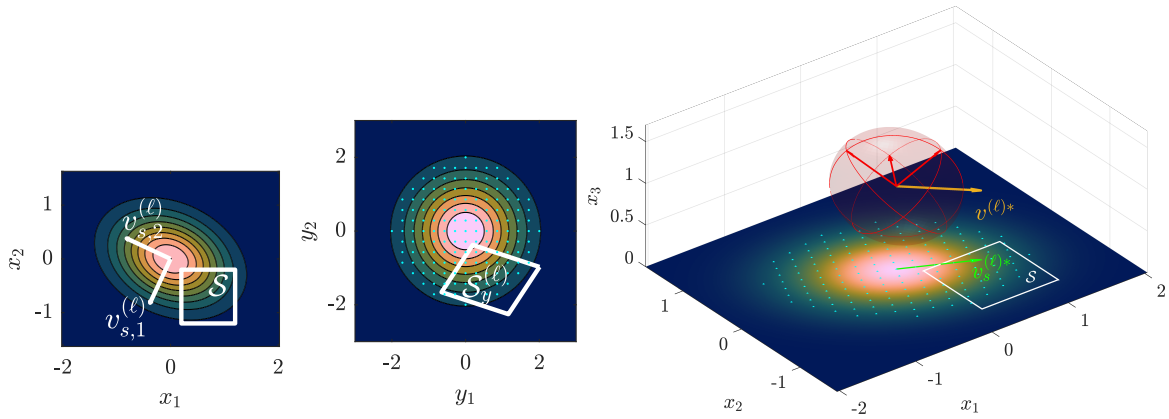


Figure 2: Original mixand position marginal (left), collocation points and transformed FoV in whitened coordinates (center), and optimal non-principal full-state split direction (right).

in the belief by way of the nonlinear state-dependent probability of detection in (22) and an expansion approximation about the mixand means. The accuracy of this approximation depends on the resolution of the mixture around the FoV boundaries $\partial\mathcal{S}$, and more precisely, about the boundaries $\partial(\mathcal{S}\cap\mathcal{V})$, where $\mathcal{V}$ represents the set of illuminated and non-occluded target positions. For ease of exposition, this intersection region will be referred to as $\mathcal{S}$ hereon. Variations in probability of detection within a GM component's local support, especially for components whose position-marginal density overlaps with sensor FoV boundaries potentially causing poor approximation and filter divergence. To address this issue, an efficient algorithm for recursively splitting a GM about the boundaries of an n-dimensional FoV is presented in [35, 36]. This paper adopts the original splitting algorithm of [35, 36] with a modification to allow splitting to occur in any direction rather than just along the principal axes of the covariance matrix.

Through a whitening transformation and collocation point inclusion test, the splitting algorithm chooses the best position-marginal covariance principal axis $\boldsymbol{v}_s^{(\ell)*}$, as shown in Figure 2. This paper proposes an improved method for selecting a splitting direction in the full-dimensional space $\mathbb{X}$. The problem is formulated as the constrained optimization problem

$$\boldsymbol{v}^{(\ell)*} \propto \underset{\boldsymbol{x}^\top (P^{(\ell)})^{-1}\boldsymbol{x}=1}{\arg\max} \left(\begin{bmatrix} (\boldsymbol{v}_s^{(\ell)})^\top & 0 \end{bmatrix} \boldsymbol{x} \right) \tag{23}$$

where the constraint in (23) accounts for the potentially non-isotropic mixand variance by limiting directions to points on the 1σ ellipsoid. The optimal split direction is then given by the unit vector

$$\boldsymbol{v}^{(\ell)*} = P^{(\ell)} \begin{bmatrix} (\boldsymbol{v}_s^{(\ell)*})^\top & 0 \end{bmatrix}^\top \tag{24}$$

Gaussian splitting is then performed along this generally non-axis-aligned direction according to [37], which generalizes previous directional split methods [38, 39].

Reward Computation The generative model computes information-theoretic reward for action-observation pairs, computation of which is now described. From (22), the posterior is

$$p_k(\boldsymbol{x}_k|Z_{1:k}) = \tfrac{1}{c} L_Z(\boldsymbol{x}_k) p_{k|k-1}(\boldsymbol{x}_k|Z_{1:k-1}) \tag{25}$$

where $L_Z(\boldsymbol{x}_k) = [\delta_\emptyset(Z_k)(1-p_{D,k}(\boldsymbol{x}_k;\boldsymbol{\xi}_k)) + (1-\delta_\emptyset(Z_k))p_{D,k}(\boldsymbol{x}_k;\boldsymbol{\xi}_k)p_k(\boldsymbol{z}_k|\boldsymbol{x}_k)]$ is the pseudo-likelihood function and c is the normalization constant. The KLD in (14) can be written as

$$D_{\mathrm{KL}}(p_k\|p_{k|k-1}) = \frac{1}{c}\int_{\mathbb{X}} p_{k|k-1}(\boldsymbol{x}_k) L_Z(\boldsymbol{x}_k) \log\left(\frac{L_Z(\boldsymbol{x}_k)}{c}\right) d\boldsymbol{x}_k \tag{26}$$

The arguments of p_k and $p_{k|k-1}$ are suppressed for brevity. In the case of a null detection ($Z_k = \emptyset$), the above KLD expression reduces to

$$\begin{aligned} D_{\mathrm{KL}}(p_k(\cdot|Z_k = \emptyset)\|p_{k|k-1}) =& \frac{1}{c_\emptyset} \int_{\mathcal{S}_k \times \mathbb{X}_v} p_{k|k-1}(\cdot)(1 - p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k)) \log(1 - p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k)) d\boldsymbol{x}_k \\ & - \frac{\log(c_\emptyset)}{c_\emptyset} + \frac{\log(c_\emptyset)}{c_\emptyset} \int_{\mathcal{S}_k \times \mathbb{X}_v} p_{k|k-1}(\boldsymbol{x}_k) p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) d\boldsymbol{x}_k \end{aligned} \tag{27}$$

where $c_\emptyset = 1 - \int_{\mathcal{S}_k \times \mathbb{X}_v} p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) p_{k|k-1}(\boldsymbol{x}_k) d\boldsymbol{x}_k$ is the normalization constant for the null detection case. Furthermore, if the probability of detection is homogeneous within $\mathcal{S}_k$ such that $p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) = 1_{\mathcal{S}_k}(\boldsymbol{x}_k) \cdot p_{D,k} \; \forall \, \boldsymbol{r}_k \in \mathcal{S}_k$, then $c_\emptyset = 1 - p_{D,k} \lambda_{\mathcal{S}_k}$ and

$$D_{\mathrm{KL}}(p_k(\cdot|Z_k = \emptyset)\|p_{k|k-1}) = \left(\frac{1 - p_{D,k}}{1 - p_{D,k} \cdot \lambda_{\mathcal{S}_k}} \right) \cdot \lambda_{\mathcal{S}_k} \cdot \log(1 - p_{D,k}) - \log(1 - p_{D,k} \lambda_{\mathcal{S}_k}) \tag{28}$$

where $\lambda_{\mathcal{S}_k} = \int_{\mathcal{S}_k} p_{k|k-1}(\boldsymbol{r}_k) d\boldsymbol{r}_k$ is the predicted cardinality of objects in $\mathcal{S}_k$.

In the singleton measurement case ($Z_k = \{\boldsymbol{z}_k\}$), the normalization constant c is

$$c = \int_{\mathcal{S} \times \mathbb{X}_v} p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) p_k(\boldsymbol{z}_k|\boldsymbol{x}_k) p_{k|k-1}(\boldsymbol{x}_k) d\boldsymbol{x}_k \tag{29}$$

and the KLD is given as

$$\begin{aligned} D_{\mathrm{KL}}(p_k(\cdot|Z_k = \{\boldsymbol{z}_k\})\|p_{k|k-1}) =& \frac{1}{c} \int_{\mathcal{S}_k \times \mathbb{X}_v} p_{k|k-1}(\cdot) p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) \\ & \times p_k(\boldsymbol{z}_k|\boldsymbol{x}_k) \log \left(\frac{p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) p_k(\boldsymbol{z}_k|\boldsymbol{x}_k)}{c} \right) d\boldsymbol{x}_k \end{aligned} \tag{30}$$

Let $\bar{p}(\boldsymbol{x}) \triangleq \frac{p_{k|k-1}(\cdot) 1_{\mathcal{S}_k}(\boldsymbol{x})}{\lambda_{\mathcal{S}_k}}$, the normalization constant can then be computed using Monte Carlo integration as

$$c = \lambda_{\mathcal{S}_k} \int_{\mathcal{S}_k \times \mathbb{X}_v} p_{D,k}(\boldsymbol{x}; \boldsymbol{\xi}_k) p(\boldsymbol{z}_k|\boldsymbol{x}_k) \bar{p}(\boldsymbol{x}_k) d\boldsymbol{x}_k \approx \frac{\lambda_{\mathcal{S}_k}}{N} \sum_{i=1}^{N} p_{D,k}(\boldsymbol{x}_k^i; \boldsymbol{\xi}_k) p(\boldsymbol{z}_k|\boldsymbol{x}_k^i) \tag{31}$$

where $\boldsymbol{x}_k^i \sim \bar{p}(\boldsymbol{x}_k)$ are samples drawn from prior distribution within the FoV and N is the number of samples. Using (31), the KLD in (30) can be similarly approximated as

$$D_{\mathrm{KL}}(p_k(\cdot|Z_k = \{\boldsymbol{z}_k\})\|p_{k|k-1}) \approx \frac{\lambda_{\mathcal{S}_k}}{N \cdot c} \sum_{i=1}^{N} p_{D,k}(\boldsymbol{x}_k^i; \boldsymbol{\xi}_k) p(\boldsymbol{z}_k|\boldsymbol{x}_k^i) \log \left(\frac{p_{D,k}(\boldsymbol{x}_k^i; \boldsymbol{\xi}_k) p(\boldsymbol{z}_k|\boldsymbol{x}_k^i)}{c} \right) \tag{32}$$

The method is particularly efficient as it directly estimates the KLD without requiring explicit computation of the full posterior distribution.

3.3 Observation Likelihood Computation

Unlike traditional MCTS methods, the proposed approach leverages known probabilistic models to curtail the high-dimensional action and observation space. In particular, observation widening enables the reuse of generated observation nodes in frequency proportional to their likelihood. Because of the high rate at which such likelihood evaluations are required, an efficient approach to doing so is presented in this section. Online sampling requires repeated evaluation of the observation likelihood conditioned on a particular action sequence.

The likelihood can be computed as the predicted measurement density (conditioned on the action sequence giving rise to $\mathcal{S}_k$) at t_k which is given as

$$l_k(Z_k) = \int f_k(Z_k \mid \boldsymbol{x}_k; \boldsymbol{\xi}_k) p_{k|k-1}(\boldsymbol{x}_k|Z_{1:k-1}) d\boldsymbol{x}_k \tag{33}$$

where $f_k(\cdot)$ is the state conditioned measurement likelihood as defined in (12).

In the case of an empty or null measurement $Z_k = \emptyset$, the likelihood then becomes

$$l(Z_k = \emptyset) = 1 - \int p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) p_{k|k-1}(\boldsymbol{x}_k | Z_{1:k-1}) d\boldsymbol{x}_k \tag{34}$$

The latter term is equal to the *expected number of target-originated measurements* and needs to be computed only once for each possible FoV.

In the case that Z is non-empty,

$$l_k(Z_k = \{\boldsymbol{z}\}) = \int p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) p_k(\boldsymbol{z}_k | \boldsymbol{x}_k) p_{k|k-1}(\boldsymbol{x}_k | Z_{1:k-1}) d\boldsymbol{x}_k \tag{35}$$

This expression can be evaluated using Monte Carlo integration similar to (32). If the probability of detection within the FoV is constant, then $p_{D,k}(\boldsymbol{x}_k; \boldsymbol{\xi}_k) = 1_{\mathcal{S}_k}(\boldsymbol{x}_k) \cdot p_{D,k}$ where $p_{D,k}$ is a constant for a FoV. Then,

$$l_k(Z_k = \{\boldsymbol{z}_k\}) = p_{D,k} \cdot \int 1_{\mathcal{S}_k}(\boldsymbol{x}_k) p_k(\boldsymbol{z}_k | \boldsymbol{x}_k) p_{k|k-1}(\boldsymbol{x}_k | Z_{1:k-1}) d\boldsymbol{x}_k \tag{36}$$

The prior $p_{k|k-1}(\boldsymbol{x}_k | Z_{1:k-1})$ is a GM given as

$$p_{k|k-1}(\boldsymbol{x}_k | Z_{1:k-1}) = \sum_{\ell=1}^{L_{k|k-1}} w_{k|k-1}^{(\ell)} \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{m}_{k|k-1}^{(\ell)}, P_{k|k-1}^{(\ell)}) \tag{37}$$

where $L_{k|k-1}$ is the number of mixture components. Then, (36) can be approximated as

$$l_k(Z_k = \{\boldsymbol{z}_k\}) = p_{D,k} \cdot \int 1_{\mathcal{S}_k}(\boldsymbol{x}_k) \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{h}(\boldsymbol{x}_k), R_k) \sum_{\ell=1}^{L_{k|k-1}} w_{k|k-1}^{(\ell)} \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{m}_{k|k-1}^{(\ell)}, P_{k|k-1}^{(\ell)}) d\boldsymbol{x}_k \tag{38}$$

$$\approx p_{D,k} \cdot \int 1_{\mathcal{S}_k}(\boldsymbol{x}_k) \sum_{\ell=1}^{L_{k|k-1}} q^{(\ell)}(\boldsymbol{z}_k) w_{k|k-1}^{(\ell)} \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{m}_{+}^{(\ell)}, P_{+}^{(\ell)}) d\boldsymbol{x} \tag{39}$$

where, through linearization about the mixand means,

$$\boldsymbol{m}_{+}^{(\ell)} = \boldsymbol{m}_{k|k-1}^{(\ell)} + K_k^{(\ell)}(\boldsymbol{z}_k - \boldsymbol{h}(\boldsymbol{m}_{k|k-1}^{(\ell)})) \tag{40}$$

$$P_{+}^{(\ell)} = P_{k|k-1}^{(\ell)} - K_k^{(\ell)} H^{(\ell)} P_{k|k-1}^{(\ell)} \tag{41}$$

$$K_k^{(\ell)} = P_{k|k-1}^{(\ell)} H^{(\ell)\top} (H^{(\ell)} P_{k|k-1}^{(\ell)} H^{(\ell)\top} + R_k)^{-1} \tag{42}$$

$$H^{(\ell)} = \frac{d}{d\boldsymbol{x}_k} \boldsymbol{h}(\boldsymbol{x}_k) \Big|_{\boldsymbol{x}_k = \boldsymbol{m}_{k|k-1}^{(\ell)}} \tag{43}$$

$$q^{(\ell)}(\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{h}(\boldsymbol{m}_{k|k-1}^{(\ell)}), H^{(\ell)} P_{k|k-1}^{(\ell)} H^{(\ell)\top} + R_k) \tag{44}$$

To obtain a closed-form density expression over $\boldsymbol{z}_k$ that can be evaluated quickly for any sample $\boldsymbol{z}_k$, the approximation

$$\int_{\mathcal{S} \times \mathbb{X}_v} \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{m}^{+}, P^{+}) d\boldsymbol{x} \approx 1_{\mathcal{S}_k}(\boldsymbol{m}_{+}) \tag{45}$$

is made, which is valid when the updated mixand $\boldsymbol{m}^{+}, P^{+}$ is mostly contained within the FoV $\mathcal{S}_k$.

Let $\mathcal{S}_k' = \{\boldsymbol{h}(\boldsymbol{x}); \boldsymbol{x} \in \mathcal{S}_k\} \subset \mathbb{R}^2$ be the measurement space image of $\mathcal{S}_k$ under $\boldsymbol{h}(\cdot)$, and $1_{\mathcal{S}_k'}(\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x})) = 1_{\mathcal{S}_k}(\boldsymbol{x})$ be indicator function for $\mathcal{S}_k'$. Using the aforementioned approximation,

$$\int 1_{\mathcal{S}_k}(\boldsymbol{x}_k) w_{k|k-1}^{(\ell)} \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{m}_{+}^{(\ell)}, P_{+}^{(\ell)}) d\boldsymbol{x}_k \approx w_{k|k-1}^{(\ell)} 1_{\mathcal{S}_k'}(\boldsymbol{z}_k) 1_{\mathcal{S}_k}(\boldsymbol{m}_{+}^{(\ell)}) \tag{46}$$

for a measurement sample $\boldsymbol{z}_k$. Thus

$$l(Z = \{\boldsymbol{z}_k\}) = p_{D,k} \cdot 1_{\mathcal{S}'_k}(\boldsymbol{z}_k) \sum_{\ell=1}^{L_{k|k-1}} q^{(\ell)}(\boldsymbol{z}_k) w_{k|k-1}^{(\ell)} 1_{\mathcal{S}_k}(\boldsymbol{m}_+^{(\ell)}) \tag{47}$$

4 Tracking

The MCTS-based controller works in conjunction with an adaptive Gaussian mixture interacting multiple model (AGMIMM) tracker [40, 41] to maintain estimates of the CSO state. The AGMIMM additionally estimates the CSO modality to account for unknown maneuvers. For simplicity, the MCTS planner operates under a ballistic-only assumption, which is reasonable over short planning horizons. The control policy in MCTS is formulated on the basis of kinematic state uncertainty reduction and thus the controller information states are marginals of the tracker's joint distribution.

The CSO state uncertainty evolves into complex non-Gaussian distributions over time due to the highly non-linear and chaotic nature of the CR3BP dynamics. Moreover, the the measurement function (10) is nonlinear, which can cause the posterior distribution to become significantly non-Gaussian after measurement updates. AGMIMM employs mixture splitting during both prediction and update stages to handle such nonlinearities. The splitting direction is determined using an uncertainty-scaled second-order linearization change method [37], which leverages the continuous-time second-order partial derivative tensor of the dynamics as an efficient surrogate for the state transition tensor over short time steps. To maintain computational efficiency, the belief updates in MCTS skip most nonlinearity-based splitting operations, retaining only the essential FoV-based splitting needed for properly constructing multi-stage branches with null detections.

5 Results

To evaluate the performance of the proposed MCTS algorithm, the reacquisition of the Artemis I-like CSO following its pass behind the Moon is considered. A simplified mission trajectory is generated as described in [41]. The CSO is considered lost after its last observation prior to the lunar flyby. Based on the predicted ballistic trajectory, the reacquisition operation is initiated on November 22, 2022 at 01:36:51 UTC, when the CSO is expected to emerge post the lunar flyby. The search is conducted over a six-hour period following this initial epoch. The CSO state distribution at this epoch, derived from trajectory propagation, serves as the *a priori* information to guide the search. A single space-based observer is employed to reacquire the target. The state of the observer in the synodic frame at the start of the search is

$$\boldsymbol{r}_{o,0}^{\top} = \begin{bmatrix} 5.23766 \times 10^5 & 1.24323 \times 10^5 & 0 \end{bmatrix} \text{ km} \tag{48}$$

$$\boldsymbol{v}_{o,0}^{\top} = \begin{bmatrix} 0.3065 & -0.7927 & 0 \end{bmatrix} \text{ km/s} \tag{49}$$

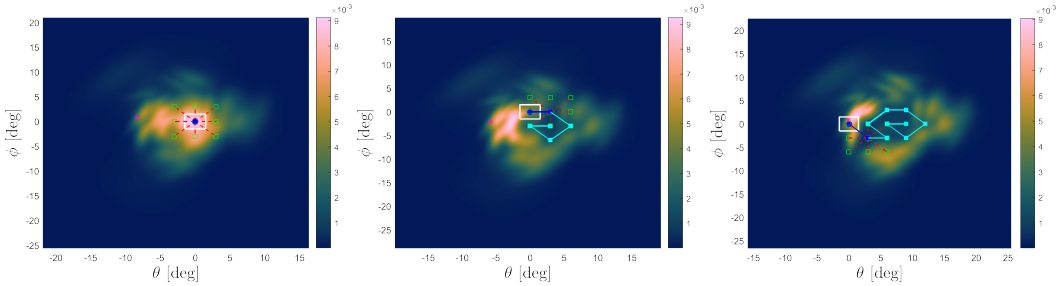
The square pyramidal sensor used has an angular width of 3° and a maximum range of two times the lunar orbit semi-major axis. Due to constraints on the observer's motion, the sensor's FoV is limited to a maximum slew of $3\sqrt{2}^\circ$ between consecutive time steps. The sensor generates clutter-free measurements when the CSO is within the FoV, outside solar exclusion, illuminated, and not occluded by the Moon. Figure 3a shows the initial distribution in measurement space centered around the initial sensor FoV (projected onto the lower dimensional measurement space) shown as a white box. The possible sensor pointing directions (depicted as green square markers) form the action space at this instant, while the CSO is marked by a magenta star.

The primary tuning parameters for MCTS include the search depth d which determines the planning horizon and the number of iterations N_{iter} used to construct the tree. The observation progressive widening parameters are set to $k_o = 6$ and $\alpha_o = 0.15$. These values are selected to complement the finite cardinality of the admissible action set while ensuring an appropriate balance in tree widening and available computational resources. Specifically, these parameters allow sufficient initial exploration through frequent node expansion when visit counts are low, while progressively becoming more selective at higher iteration counts.

Unless explicitly stated otherwise, the probability mass-based greedy rollout policy is employed as the default configuration due to its ability to account for the probability of detection during the search. The

iteration count is critical for ensuring adequate exploration of the action space through sufficient visit counts across different actions. Insufficient iterations can lead to premature convergence and biased search behavior due to inadequate exploration of the action space. However, a key computational efficiency is achieved through the caching of observation nodes corresponding to null detection observations. Since null detections are the predominant observation type and are solely determined by the sensor's FoV, their corresponding observation nodes can be cached and reused across iterations. Consequently, the computational complexity of the algorithm scales largely with the size of the action space rather than the number of iterations, making it computationally tractable even with large iteration counts necessary for thorough exploration.

Figure 3 depicts the successful reacquisition of the CSO using an MCTS-based information-driven control policy. The evolution of the sensor's FoV under the policy is shown in Figures 3b and 3c, illustrating an intermediate search step and the step when target detection occurs, respectively. The blue rectangular boundaries trace the chronological progression of the sensor's FoV, providing visualization of the search policy's spatial coverage and the sequence of sensing actions executed by the controller.



(a) Initial distribution and FoV (b) Distribution after 24 minutes (c) Distribution after 40 minutes

Figure 3: CSO reacquisition search, showing the current FoV, the considered actions, and the past actions overlaid on the measurement probability density.

The proposed MCTS-based approach for CSO reacquisition is evaluated against two baseline methods: an ϵ-greedy search and a pure random search. The ϵ-greedy search algorithm selects actions that maximize the expected probability mass within the sensor's FoV with probability $1 - \epsilon$, and a random action with probability ϵ, where $\epsilon \in [0, 1]$ controls the explore-exploit trade-off. As ϵ increases, the algorithm encourages random exploration. This probability mass-based greedy criterion is chosen because it directly relates to the probability of target discovery within the sensor's FoV. Different ϵ values of $0.70, 0.32$ and 0.10 are compared. The random search strategy executes by uniformly sampling actions from the admissible action space at each decision epoch. This sampling is performed without regard to the underlying belief state providing a baseline stochastic search policy against which to evaluate more sophisticated strategies.

The performance evaluation for different search strategies is conducted through 100 independent Monte Carlo trials that begin with identical initial state distribution and identical true target trajectory. This allows evaluating and comparing each algorithm's performance under different stochastic realizations of the search process while maintaining consistent ground truth conditions. The performance is compared using time-to-detection as the primary metric. Trials in which the CSO is not detected are assigned the maximum time-to-detection value of six hours.

Figure 4 demonstrates the comparison of target reacquisition times for different configurations of the baseline methods and MCTS implementations with a depth of two and varying number of iterations. For the baseline random search policy, only 25% of the trials detected the CSO within the search duration. Because the trials that fail to find the CSO are assigned the maximum value of six hours, the first quartile, median, and third quartile are all six hours, so the corresponding box is compressed. Even among the trials successful in reacquisition, there is a significant lack of consistency in the detection times, demonstrating its ineffectiveness for cislunar target reacquisition. This poor performance of the random search at even finding the

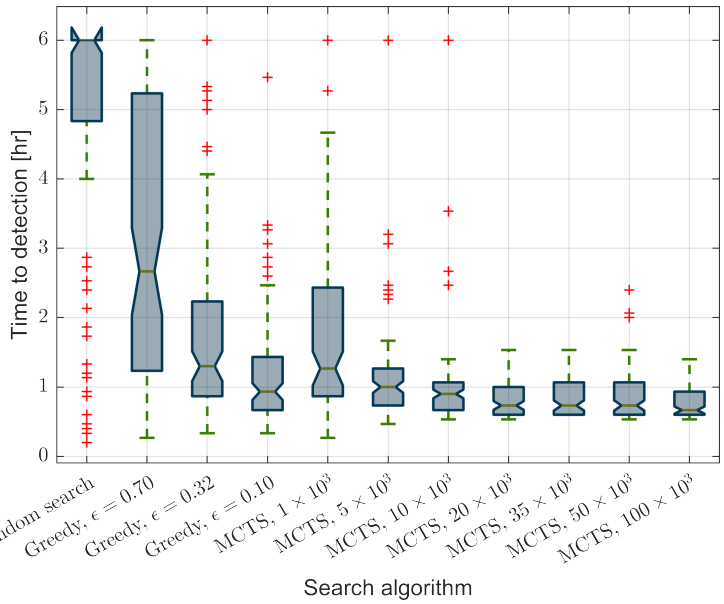


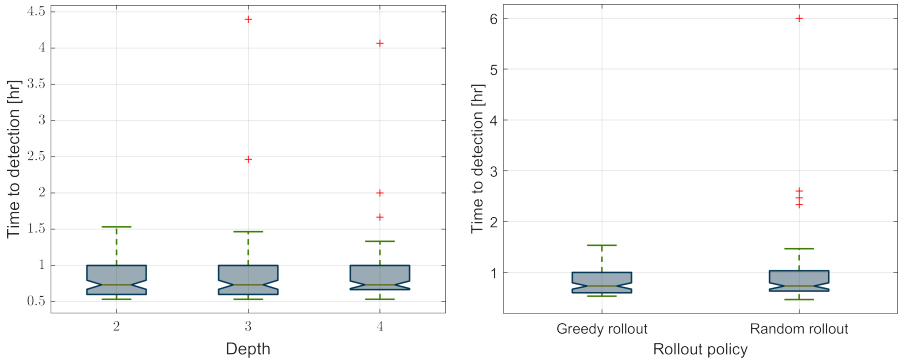
Figure 4: Time until CSO reacquisition for different search strategies.

CSO–much less in a timely manner–clearly demonstrates the need for more intelligent searching schemes in such challenging scenarios.

The ϵ-greedy approach shows improvement in median time-to-detection as the exploration parameter ϵ or equivalently the randomness in action selection decreases. This suggests that exploitation of the probability mass-heuristic is at least reliable for a myopic target reacquisition. The advantage of MCTS over traditional ϵ-greedy algorithm, even with only one additional depth, is evident from Figure 4. MCTS with 1,000-5,000 iterations demonstrates suboptimal performance due to insufficient exploration of the action-observation space. However, increasing the iteration count yields significant performance improvements, especially over $10,000$ iterations, characterized by both reduced median time-to-detection and lower performance variance compared to all ϵ-greedy configurations. The reduction in variance and outlier cases suggests that the increased iteration count enables MCTS to thoroughly exhaust the action-observation space and compute a better estimate of the optimal value function, resulting in robust and consistent search strategies. Beyond 10,000 iterations, each configuration managed to detect the CSO within the allotted time frame in every trial. However, the returns diminish beyond 20,000. Increasing iterations to 100,000, though not significantly computationally expensive, only offers a marginal improvement of around 7% in the detection time. This suggests that moderate iteration counts (20,000-35,000) are sufficient to explore the action-observation space for a depth of two for this particular setup.

The impact of the MCTS depth as well as that of greedy and random rollout policies on the time to detection is also investigated. Figure 5a demonstrates that increasing the planning horizon (MCTS depth) yields only marginal improvements in median time-to-detection though it does decrease the performance variance. Notably, increasing the number of iterations at fixed depth produces substantially greater performance improvements. This asymmetric impact of parameters is particularly relevant from a computational perspective as increasing the number of iterations (even by an order of magnitude) imposes much less computational strain than increasing the search depth by one level.

The comparison of rollout policies in Figure 5b reveals that the probability mass-based greedy rollout achieves comparable performance to random rollouts, with the primary difference being increased robustness against outlier cases and search failures. Given the comparable performance between the two rollout strategies, the additional computational cost associated with the probability mass-based greedy rollout policy may not be justified, especially in computationally constrained scenarios, where random rollout may suffice. The



(a) MCTS, 20,000 iterations, varying depths

(b) MCTS, 20,000 iterations, varying rollout policies

Figure 5: Time until CSO reacquisition for MCTS with different depths and rollout policies.

modest performance advantage of greedy rollouts over random rollouts suggest that while probability mass-based heuristic is suitable for immediate action selection, it may not be particularly informative for evaluating long-term action sequences. The relationship between probability mass maximization and optimal long-term search behavior is weaker than implicitly assumed in this paper.

6 Conclusion

An information-driven approach to space-based sensor motion planning for reacquiring a lost object in cislunar space is presented. The problem is formulated as a stochastic optimal control problem and MCTS with DPW is used to obtain non-myopic search policies. The results demonstrate the robustness of the MCTS-based approach for CSO reacquisition while effectively handling the chaotic nature of the cislunar orbital motion, the stochasticity of the target-motion and sensing models, constraints in observer motion, and the prevalence of null observations during the search process. Through extensive Monte Carlo analysis, MCTS with sufficient iterations is shown to outperform both random and ϵ-greedy strategies, achieving lower median time-to-detection and reduced performance variance. MCTS demonstrates superior performance, particularly in scenarios where the true target state lies in low-probability regions of the belief state, distant from both the maximum a posteriori and maximum likelihood estimates of the state distribution. However, the performance of the probability mass-based greedy rollouts in MCTS is not significantly better than that of the random rollouts, despite the additional computational overhead required by the former. This indicates that the current greedy rollout might not be effective at estimating the long term reward, leaving room for more sophisticated rollout policies that can potentially enhance the search.

7 Acknowledgment

Part of this work was sponsored by the United States Air Force Research Laboratory and the United States AFRL Regional Hub and was accomplished under Cooperative Agreement Number FA8750-22-2-0501. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] E. M. Gaposchkin, C. Von Braun, and J. Sharma, "Space-Based Space Surveillance with the Space-Based Visible," *Journal of Guidance, Control, and Dynamics*, Vol. 23, No. 1, 2000, pp. 148–152.

[2] R. P. Mahler, *Advances in Statistical Multisource-Multitarget Information Fusion*. Artech House, 2014.

[3] P. S. Williams, D. B. Spencer, and R. S. Erwin, "Coupling of Estimation and Sensor Tasking Applied to Satellite Tracking," *Journal of Guidance, Control, and Dynamics*, Vol. 36, No. 4, 2013, pp. 993–1007.

[4] M. Patel, A. J. Sinclair, and K. Ho, "Information-Theoretic Target Search for Space Situational Awareness," *2018 Space Flight Mechanics Meeting*, 2018, p. 0725.

[5] N. Adurthi, P. Singla, and M. Majji, "Mutual Information Based Sensor Tasking with Applications to Space Situational Awareness," *Journal of Guidance, Control, and Dynamics*, Vol. 43, No. 4, 2020, pp. 767–789.

[6] K. A. LeGrand, P. Zhu, and S. Ferrari, "A Random Finite Set Sensor Control Approach for Vision-based Multi-object Search-While-Tracking," *2021 24th International Conference on Information Fusion (FUSION)*, 2021.

[7] K. A. LeGrand, P. Zhu, and S. Ferrari, "Cell Multi-Bernoulli (Cell-MB) Sensor Control for Multi-Object Search-While-Tracking (SWT)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, June 2023, pp. 7195–7207, 10.1109/TPAMI.2022.3223856.

[8] J. Africano, T. Schildknecht, M. Matney, P. Kervin, E. Stansbery, and W. Flury, "A Geosynchronous Orbit Search Strategy," *Space Debris*, Vol. 2, 2000, pp. 357–369.

[9] T. Schildknecht, "Optical Surveys for Space Debris," *The Astronomy and Astrophysics Review*, Vol. 14, 2007, pp. 41–111.

[10] S. N. Paul, B. D. Little, and C. Frueh, "Detection of Unknown Space Objects Based on Optimal Sensor Tasking and Hypothesis Surfaces Using Variational Equations," *The Journal of the Astronautical Sciences*, Vol. 69, No. 4, 2022, pp. 1179–1215.

[11] S. Gehly, B. Jones, and P. Axelrad, "Sensor Allocation for Tracking Geosynchronous Space Objects," *Journal of Guidance, Control, and Dynamics*, Vol. 41, No. 1, 2018, pp. 149–163.

[12] C. Frueh, H. Fielder, and J. Herzog, "Heuristic and Optimized Sensor Tasking Observation Strategies with Exemplification for Geosynchronous Objects," *Journal of Guidance, Control, and Dynamics*, Vol. 41, No. 5, 2018, pp. 1036–1048.

[13] B. D. Little and C. E. Frueh, "Space Situational Awareness Sensor Tasking: Comparison of Machine Learning with Classical Optimization Methods," *Journal of Guidance, Control, and Dynamics*, Vol. 43, No. 2, 2020, pp. 262–273.

[14] L. Federici, A. D'Ambrosio, R. Furfaro, and V. Reddy, "Optimal Sensor Tasking for Space Domain Awareness via a Beam A*-Search Algorithm," *Proceedings of the Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*, 2023, p. 56.

[15] N. Adurthi, P. Singla, and M. Majji, "Dynamic Data-Driven Sensor Tasking with Applications in Space and Aerospace Systems," *Handbook of Dynamic Data Driven Applications Systems: Volume 2*, pp. 249–283, Springer, 2023.

[16] H. Cai, Y. Yang, S. Gehly, C. He, and M. Jah, "Sensor Tasking for Search and Catalog Maintenance of Geosynchronous Space Objects," *Acta Astronautica*, Vol. 175, 2020, pp. 234–248.

[17] S. Fedeler and M. Holzinger, "Monte Carlo Tree Search Methods for Telescope Tasking," *AIAA Scitech 2020 Forum*, 2020, p. 0659.

[18] S. J. Fedeler, M. J. Holzinger, and W. W. Whitacre, "Tasking and Estimation for Minimum-Time Space Object Search and Recovery," *The Journal of the Astronautical Sciences*, Vol. 69, No. 4, 2022, pp. 1216–1249.

[19] R. Linares and R. Furfaro, "An Autonomous Sensor Tasking Approach for Large Scale Space Object Cataloging," *Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS)*, 2017, pp. 1–17.

[20] P. M. Siew, D. Jang, and R. Linares, "Sensor Tasking for Space Situational Awareness Using Deep Reinforcement Learning," *Proceedings of the AAS/AIAA Astrodynamics Specialist Conference, Big Sky, MT, USA*, 2021, pp. 9–11.

[21] K. Tomita, Y. Shimane, and K. Ho, "Multi-Spacecraft Predictive Sensor Tasking for Cislunar Space Situational Awareness," *arXiv preprint arXiv:2310.04894*, 2023.

[22] S. Fedeler, M. Holzinger, and W. Whitacre, "Sensor Tasking in the Cislunar Regime Using Monte Carlo Tree Search," *Advances in Space Research*, Vol. 70, No. 3, 2022, pp. 792–811.

[23] P. M. Siew, D. Jang, T. G. Roberts, R. Linares, and J. Fletcher, "Cislunar Space Situational Awareness Sensor Tasking Using Deep Reinforcement Learning Agents," *2022 Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS), Maui, Hawaii*, 2022.

[24] R. T. Eapen, S. N. Paul, and P. Singla, "Sensor Tasking Strategies for Space-Based Observers in the Cislunar Environment," *AIAA SCITECH 2024 Forum*, 2024, p. 1676.

[25] M. J. Holzinger, D. J. Scheeres, and K. T. Alfriend, "Object Correlation, Maneuver Detection, and Characterization Using Control Distance Metrics," *Journal of Guidance, Control, and Dynamics*, Vol. 35, No. 4, 2012, pp. 1312–1325.

[26] D. P. Lubey and D. J. Scheeres, "Identifying and Estimating Mismodeled Dynamics via Optimal Control Policies and Distance Metrics," *Journal of Guidance, Control, and Dynamics*, Vol. 37, No. 5, 2014, pp. 1512–1523.

[27] N. Singh, J. T. Horwood, and A. B. Poore, "Space Object Maneuver Detection via a Joint Optimal Control and Multiple Hypothesis Tracking Approach," *Proceedings of the 22nd AAS/AIAA Space Flight Mechanics Meeting*, Vol. 143, Univelt San Diego, CA, 2012, pp. 843–862.

[28] G. M. Goff, D. Showalter, J. T. Black, and J. A. Beck, "Parameter Requirements for Noncooperative Satellite Maneuver Reconstruction Using Adaptive Filters," *Journal of Guidance, Control, and Dynamics*, Vol. 38, No. 3, 2015, pp. 361–374.

[29] Z. Hall, D. Schwab, R. Eapen, and P. Singla, "Reachability-Based Approach for Search and Detection of Maneuvering Cislunar Objects," *AIAA SCITECH 2022 Forum*, 2022, p. 0853.

[30] Z. Sunberg and M. Kochenderfer, "Online Algorithms for POMDPs with Continuous State, Action, and Observation Spaces," *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 28, 2018, pp. 259–263.

[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, New York [etc.]: J. Wiley & sons, 1991.

[32] D. Alspach and H. Sorenson, "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," *IEEE transactions on automatic control*, Vol. 17, No. 4, 1972, pp. 439–448.

[33] A. Couëtoux, J.-B. Hoock, N. Sokolovska, O. Teytaud, and N. Bonnard, "Continuous Upper Confidence Trees," *Learning and Intelligent Optimization: 5th International Conference, LION 5, Rome, Italy, January 17-21, 2011. Selected Papers 5*, Springer, 2011, pp. 433–445.

[34] K. A. LeGrand, A. V. Khilnani, and J. Iannamorelli, "Bayesian Angles-Only Cislunar Space Object Tracking," *33rd AAS/AIAA Space Flight Mechanics Meeting*, 2023.

[35] K. LeGrand and S. Ferrari, "The Role of Bounded Fields-of-View and Negative Information in Finite Set Statistics (FISST)," *Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020*, 2020, 10.23919/FUSION45008.2020.9190174.

[36] K. A. LeGrand and S. Ferrari, "Split Happens! Imprecise and Negative Information in Gaussian Mixture Random Finite Set Filtering," *Journal of Advances in Information Fusion*, Vol. 17, Dec. 2022, pp. 78–96.

[37] J. Kulik and K. A. LeGrand, "Nonlinearity and Uncertainty Informed Moment-Matching Gaussian Mixture Splitting," Nov. 2024, 10.48550/arXiv.2412.00343.

[38] V. Vittaldev and R. P. Russell, "Multidirectional Gaussian Mixture Models for Nonlinear Uncertainty Propagation,"

[39] K. Tuggle, *Model Selection for Gaussian Mixture Model Filtering and Sensor Scheduling*. PhD thesis, University of Texas at Austin, 2020.

[40] J. Iannamorelli and K. LeGrand, "Adaptive Filtering for Multi-Sensor Maneuvering Cislunar Space Object Tracking," *Proceedings of the Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*, 2023, p. 21.

[41] J. L. Iannamorelli and K. A. LeGrand, "Adaptive Gaussian Mixture Filtering for Multi-sensor Maneuvering Cislunar Space Object Tracking," *The Journal of the Astronautical Sciences*, Vol. 72, Jan. 2025, p. 2, 10.1007/s40295-024-00478-z.