Imperfect Information Games and Counterfactual Regret Minimization in Space Domain Awareness

Tyler Becker
University of Colorado Boulder
Zachary Sunberg
University of Colorado Boulder

ABSTRACT

In order to maintain space domain awareness (SDA), sensors must be assigned tasks judiciously. Previous SDA sensor tasking methods use optimization and consider stochastic model and state uncertainty. While these optimization-based approaches are robust to disturbances with known distributions, they are prone to failure if a space asset operator wants to deliberately conceal their actions or intentions or if natural disturbances have certain pathological structure. To address these shortcomings, we propose a game-theoretic approach to calculate stochastic strategies resistant to exploitation that would be impossible to determine using single-agent optimization or reinforcement learning. We study two simplified SDA examples. First, Monte Carlo counterfactual regret (MCCFR) methods are applied to a simple mode change detection game to find Nash equilibrium solution. Second, sparse MCCFR methods are applied to an orbit change game with Keplerian satellite dynamics to demonstrate convergence to unexploitable stochastic strategies for both satellite and ground station. Finally, we quantify the operational advantage of game theoretic sensor tasking, showing that it significantly reduces sensor time budget requirements in the mode change detection mission.

1. INTRODUCTION

As space becomes a more important resource, space domain awareness (SDA) will increasingly involve interaction with actors who may want to conceal their operations and intentions. Robust awareness will depend on the ability to master intelligence gathering in the highly unintuitive domain of orbital mechanics when other actors are trying to prevent it.

Consider, for instance, an operator who wants to insert a satellite into translunar orbit without being detected and a monitoring agency that wants to keep track of every object in such an orbit. Translunar injection is most efficient at certain points in an object's orbit, so the monitoring agency may want to monitor those points specifically. However, knowledge that these regions are more likely to be monitored may cause the satellite operator to adjust the maneuver location to avoid detection, and the monitor may update their schedule in response. In order to make it harder for the other player to predict and respond, it may be beneficial for these players to act with some degree of randomness (commonly called bluffing). Calculating the best strategies accounting for these factors and the complexities of orbital dynamics is extremely challenging. Fortunately, game theory provides a framework for analyzing how actors will behave and determining the best strategies to accomplish particular goals when interacting with them.

Extensive work has been conducted on sensor tasking for space domain awareness from an optimization perspective [8, 9, 6, 25, 24, 12, 13]. However, as we will make clear below, optimization is *fundamentally different* from game theoretic solutions, and there are certain useful solutions that optimization is incapable of finding. On the other hand, extensive work has also been done to solve imperfect information games like poker [4, 15, 1]. *Some* work has been done to connect the fields of game theory and SDA, but until now this has been limited to the perfect information domain [21, 20].

This paper extends imperfect information extensive form solution methods to space domain awareness, using both simple sequential models and physics-based differentiable dynamics models.

Section 2 gives a brief review of the game theoretic concepts needed to introduce the algorithms used in Section 3. These algorithms are applied to a simple sequential mode change game wherein a ground station is required to deter a satellite from undergoing a mode change in high value areas in an orbit. Then a more advanced set of algorithm variants are applied to a goal reaching game wherein a ground station is tasked with tracking and inferring the intentions of an orbiting satellite. Section 4 shows the results of these experiments.

2. BACKGROUND

2.1 Game Theory

Game theory is a field concerned with analyzing the interaction between multiple agents (also called players, each assigned an identifying number $i \in \mathcal{N}$) with possibly competing objectives. In a traditional optimization problem, there is a single objective/utility function with local optima that can be compared against each other to find a global optimum. In contrast, since each agent in a game has their own utility function (denoted u_i), the best response of each player depends upon how the other players act, so it may be impossible to find a set of behaviors that all players agree is globally optimal. Formally, each player must choose a strategy σ_i from the space of possible strategies Σ_i to maximize the expected utility resulting from the joint strategy $u_i(\sigma_i, \sigma_{-i})$, where $-i$ is the set of all players *excluding* i. The game defines utility as a function of terminal states ($z \in Z$), so we define utility of a joint strategy (σ) as the sum of terminal state utility weighted by the reach probability of that terminal state:

$$u_i(\sigma) = \sum_{z \in Z} \pi^{\sigma}(z) u_i(z). \tag{1}$$

2.2 Normal Form Games

In a Normal form game, each player plays only once. For example in the very simple two-player game of rock-paper-scissors, each player is given the option of playing either rock, paper or scissors. Paper beats rock, scissors beats paper, and rock beats scissors. The payouts for beating, tying with, or losing to an opponent are $+1$, 0, and -1, respectively.

Suppose we choose initial strategies to be pure (deterministic) with player 1 always choosing to play rock and player 2 always choosing to play paper. Furthermore, suppose each player can respond to the other's strategy by choosing a new pure strategy that is an optimal response to the other player's strategy. In this case player 1 would develop a best response that results in always playing scissors to exploit player 1's pure paper strategy. Continuing this iterative optimization would never result in a stable solution, as players would constantly iterate through all pure strategies without ending.

To reconcile this non-convergent behavior emergent from the utility maximization solution concept, we must consider more general strategies and solution concepts. In particular, in addition to pure strategies, we allow the agents to play *mixed* strategies, where the action is chosen stochastically from a distribution. The Nash equilibrium solution concept provides a coherent framework for reasoning about how players choose mixed strategies. A Nash equilibrium is a stable point for joint strategies in which neither player has any incentive to unilaterally deviate from their current strategy i.e. any deviation from a Nash equilibrium strategy will never yield greater expected utility. We define the Nash equilibrium strategy σ by

$$\forall i, u_i(\sigma) \geq \max_{\sigma_i' \in \Sigma_i} u_i(\sigma_i', \sigma_{-i}). \tag{2}$$

If mixed strategies are allowed, at least one Nash equilibrium exists for every game [16].

For the game of rock-paper-scissors, this solution criterion is met when both players play a uniformly random strategy. Because each outcome is equally likely, the expected utility for both players is zero. If one player were to deviate from this uniform strategy, their expected utility would still be zero as losing, tying, and winning are still all equally likely outcomes. Therefore, there is no incentive to unilaterally deviate. What also makes this solution concept particularly alluring is that, in zero-sum games, the Nash equilibrium solution also yields the joint strategy that maximizes worst-case expected utility [18].

2.3 Imperfect Information Extensive Form Games

While the normal form works well for modeling games where each player takes a single action, the scenarios we consider are sequential in nature and span over multiple time steps. Therefore, we use the imperfect information extensive game formalism.

In this formalism, the underlying true state of the game is defined as the history which is a sequence of actions. Each history is assigned a player whose turn it is to act. The expression $h \sqsubseteq h'$ signifies that history h is a prefix of history h'. Terminal histories $Z \subseteq H$ signify the end of the game, resulting in a payoff to each player via the utility function $U : \mathscr{N} \times Z \to \mathbb{R}$.

Because of the imperfect information nature of the game, players do not have access to history information. Rather, each player only has access to information state which is defined as a collection of plausible histories.

One method for finding Nash equilibria is regret minimization. The regret for player i and action a is the difference between the utility in taking that action and the utility realized by the current strategy, σ. For a normal-form game this can be expressed as $R_i(a) = u_i(a) - u_i(\sigma)$. At each training iteration we use a strategy that plays each action proportionally to its positive accumulated regret, i.e.

$$\sigma_i^{T+1}(a) = \frac{R_i^{T,+}(a)}{\sum_a R_i^{T,+}(a)}, \tag{3}$$

where $R^+ = \max(R, 0)$. The average of all strategies used during training, $\bar{\sigma}$, will converge to a Nash equilibrium [7].

Counterfactual regret minimization (CFR) [27] extends this approach to extensive form games by operating on each individual information set. The "counterfactual" term refers to the value calculated under the counterfactual assumption that the player used a strategy that seeks to reach that information set (I), given by

$$v_i(\sigma, I) = \sum_{z \in Z_I} \pi_{-i}^{\sigma}(z[I]) \pi^{\sigma}(z[I], z) u_i(z), \tag{4}$$

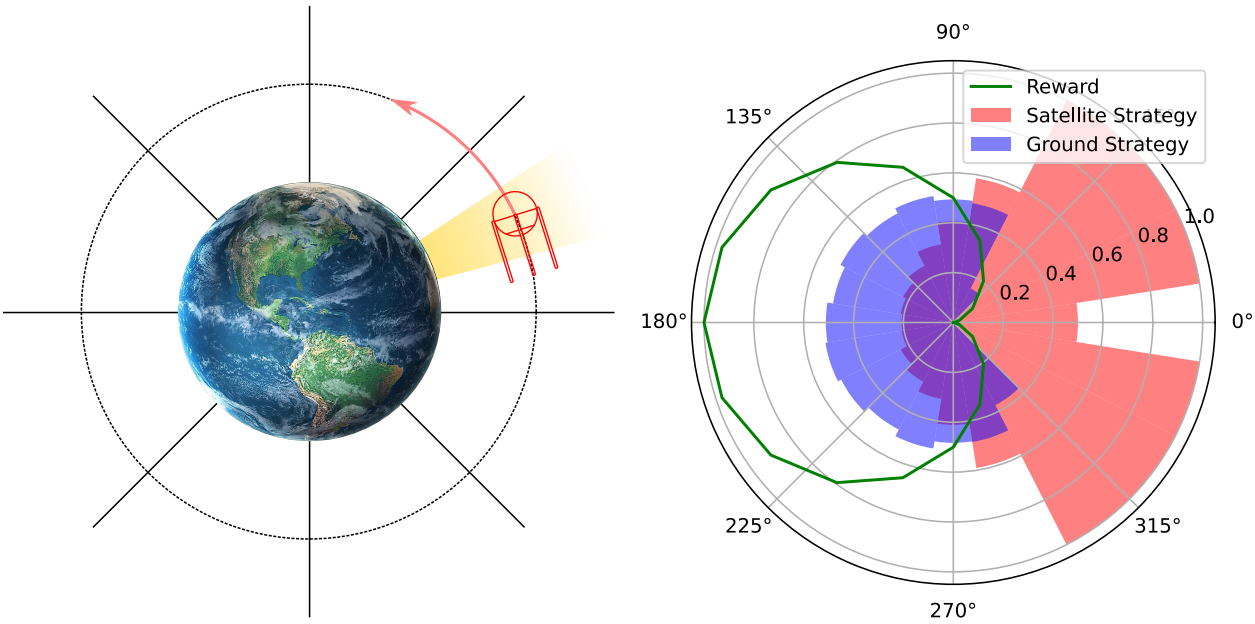
where $\pi_{-i}^{\sigma^t}(I)$ is the probability that all players except i played actions that result in reaching I. $\pi_{-i}^{\sigma^t}(I)$ is referred to as the counterfactual reach probability as it considers a scenario wherein player i didn't actually play their true strategy σ, but instead played purely to reach information set I.

The average counterfactual regret R^T for action a and information set I over T total training steps is the average of all intermediate counterfactual regrets r:

$$R_i^T(I, a) = \frac{1}{T} \sum_{t=1}^{T} \pi_{-i}^{\sigma^t}(I) \left(u_i\left(\sigma^t \big|_{I \to a}, I \right) - u_i\left(\sigma^t, I \right) \right). \tag{5}$$

Zinkevich et al. [27] show that true overall regret is upper-bounded by counterfactual regret, implying that the proposed procedure that minimizes counterfactual regret, in turn, minimizes overall regret.

However, in practice CFR is also computationally intractable in larger games due to the need to traverse the full game tree at each iteration. Variants such as CFR+ [5, 26] and discounted CFR [3] sought to improve the convergence rate per CFR iteration, decreasing the total required number of iterations. While these methods significantly decreased exploitability on a per-iteration basis relative to vanilla CFR, these methods still require traversing the full game tree at each iteration, which can be prohibitively expensive provided that the size of the game tree is exponential in depth. Lanctot et al. [10] then proposed Monte Carlo CFR (MCCFR), where instead of traversing the entire game tree at each iteration, only a randomly sampled portion of the game tree is traversed and updated. This is done in a manner that makes the Monte Carlo update the same as the vanilla update in expectation. Fortunately, CFR+ and discounted CFR can be combined with MCCFR to even further improve convergence rates. While these Monte Carlo methods helped in making larger games tractable to solve, the shortcoming was that large portions of the game tree would rarely be updated. This problem was largely ameliorated with neural network function approximation [2, 23, 14], wherein the strategy over the entire game tree would be updated despite only sparsely sampling nodes in the tree, making these function approximation methods highly sample efficient.



(a) Notional satellite orbits and ground station fields of view.

(b) Nash equilibrium strategy profile (details in Section 4.1).

Fig. 1: Mode change game.

3. EXPERIMENTAL SETUP

In order to begin evaluating game theoretic approaches for sensor tasking, we consider two example problems. The first is a mode change game wherein a satellite deterministally transitions between discrete zones around earth with the central goal of undergoing a mode change in high value zones. Here, "mode change" is used as a comprehensive term encompassing several actions the satellite could take with examples given below. The ground station's goal in this game is to detect the satellite in the act of changing modes. The second game incorporates Keplerian orbital mechanics for the satellite, where the goal of the satellite is to reach some desired orbital altitude undetected and the goal of the ground station is to detect the satellite and determine its intentions.

3.1 Mode Change Game

We begin by studying a medium-sized problem called the "mode change game" that has features representative of games that might be encountered in space domain awareness. In this game, a satellite moves through N discrete zones in an orbit (Fig. 1a). The satellite seeks to perform a "mode change" without being detected. The mode change could represent, for example, changing orbit, emitting signals, changing orientation, or deploying a payload. A network of N ground stations, each corresponding to a zone of the orbit, can scan and detect a mode change, however there is a budget representing other demands on the sensor network that limits how many times the network can scan. Mode changes are more valuable to the satellite at certain parts of the orbit. For the particular case we investigated, the reward profile is shown in Fig. 1b indicating that mode changes are most valuable at 180°.

The game size $|H| = (|A_1||A_2|)^T$ is relatively large, where T is number of discrete orbit zones, A_1 is the action space for the satellite and A_2 is the action space for the ground station. For the experiments in this paper we use $T = 20$ with binary actions spaces for both the satellite and ground station resulting in $|H| \approx 1.10 \times 10^{12}$. The information available to the ground station at any time step is the budget remaining and the current time step, and the information available to the satellite at any time step is just the current time step. Due to this limited information, the size of the information state space is relatively small ($|I| = 119$) allowing the tabular data to be easily stored on a desktop computer's random access memory. Because the full game tree would be far too computationally costly to traverse with non-sampling based CFR variants, we use external sampling CFR (ESCFR), which decreases the searched history space to $2^{20} = 1,048,576$ nodes on each iteration.

ESCFR is a form of Monte Carlo CFR (MCCFR) [10], where we no longer perform updates according to an exact

counterfactual value v, rather a sampled counterfactual value $\tilde{v}$, given by

$$\tilde{v}_i(\sigma, I \mid j) = \sum_{z \in Q_j \cap Z_I} \frac{1}{q(z)} u_i(z) \pi^{\sigma}_{-i}(z[I]) \pi^{\sigma}(z[I], z), \tag{6}$$

where Q_j is a sampled set of terminal states, and $q(z)$ is the probability of sampling terminal history z. Note that this sampled counterfactual regret is weighted such that the expectation of sampled counterfactual regret $(\tilde{r})$ is equal to the true counterfactual regret i.e. $\mathbb{E}[\tilde{r}] = r$.

For each external sampling iteration, we sample single actions from players other than the player whose strategy is currently being updated. For the updating player, all actions are still expanded in the tree traversal, yielding the following regret:

$$\tilde{r}(I, a) = (1 - \sigma(a \mid I)) \sum_{z \in Q \cap Z_I} u_i(z) \pi^{\sigma}_i(z[I]a, z), \tag{7}$$

where average counterfactual regret $\tilde{R}$ is the average of intermediate counterfactual regrets $\tilde{r}$. It is shown by Lanctot et al. [10] that external-sampling CFR requires asymptotically less time to compute an approximate equilibrium than vanilla CFR. For the mode change game, where the action space at each information state is binary, the expansion of all updating player actions is still feasible with this method. Nonetheless, were we to increase the depth of the game, external sampling could quickly become computationally intractable due to computation time still being exponential in game depth.

3.2 Orbit Change Game

For the dynamic goal reaching game, we once again task a ground station with scanning the sky for a satellite. However, the satellite's dynamics are no longer governed by a deterministic transition between sectors. Rather, the satellite's dynamics are governed by simple Keplerian orbital mechanics. Because the satellite can maneuver, the ground station must now not only determine whether or not to scan but which sector to scan.

For the paper we consider an instantiation of the game where the ground station is given 4 sectors to scan, yielding 5 total actions for the ground station: one for each sector and one to refrain from scanning altogether. For the final step of the game, however, the ground station must correctly guess the target altitude of the satellite, for which we allot four options. At the beginning of the game, the satellite is given one of 4 possible initial conditions as dealt by the chance player: four possible goal altitudes and starting at an orbit radius of $1.5R_{\text{Earth}}$. Once the chance round is complete, the satellite is allotted 3 actions: $100\frac{m}{s}$ instantaneous retrograde burn, no burn, and a $100\frac{m}{s}$ instantaneous prograde burn.

While the history space size of the goal reaching game isn't significantly larger than the mode change game for the time horizon of 10 steps that we consider ($|H| = 9.22 \times 10^{12}$), the size of the information space is orders of magnitude larger. Furthermore, the computational cost of calculating subsequent game states is significantly higher than in the mode change game due to the requirement of integrating the differential equations that govern the dynamics of the satellite at each step.

For this reason, we switch from external sampling to an even more sparse sampling method: outcome sampling. Outcome sampling, as the name suggests, samples a single outcome or terminal history by sampling from all players' strategies making the time to complete a single outcome sampling tree traversal linear in the depth of the tree. The sampled counterfactual regret for outcome sampling is given by

$$\tilde{r}(I, a) = \begin{cases} w_I \cdot (1 - \sigma(a \mid z[I])) & \text{if } (z[I]a) \sqsubseteq z \\ -w_I \cdot \sigma(a \mid z[I]) & \text{otherwise} \end{cases}, \text{ where } w_I = \frac{u_i(z) \pi^{\sigma}_{-i}(z) \pi^{\sigma}_i(z[I]a, z)}{\pi^{\sigma'}(z)}. \tag{8}$$

However, due to the sparsity of this method, the distribution of counterfactual value estimates from a single outcome sampling traversal has high variance, consequently leading to instability in large games. The proposed solution to this high variance value estimate is variance reduction via baseline subtraction, resulting in the variance reduction Monte Carlo counterfactual regret minimization algorithm (VR-MCCFR) [19]. This method introduces information state and action dependent baseline $b(\sigma, I, a)$ as a control variate.

From this, we get a new baseline subtracted utility estimate $\hat{u}_i^b$ given by

$$\hat{u}_i^b(\sigma,h,a \mid z) = \begin{cases} b_i\left(I_i(h),a\right) + \frac{\hat{u}_i^b(\sigma,ha|z) - b_i(I_i(h),a)}{\xi(h,a)} & \text{if } ha \sqsubseteq z \\ b_i\left(I_i(h),a\right) & \text{if } h \sqsubset z, ha \not\sqsubseteq z \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where $\xi(h,a)$ is the probability of sampling action a from history h given outcome sampling policy ξ.

Finally, we have the baseline subtracted counterfactual value estimate given by

$$\hat{v}_i^b(\sigma,I(h),a \mid z) = \hat{v}_i^b(\sigma,h,a \mid z) = \frac{\pi_{-i}^{\sigma}(h)}{q(h)} \hat{u}_i^b(\sigma,h,a \mid z). \tag{10}$$

Here a history value estimate made on any given iteration is bootstrapped by the result of all prior value estimates.

To further improve convergence rates, we use CFR+ [26, 5], where cumulative regret updates are clipped as follows,

$$R_i^{+,T}(I,a) = \begin{cases} \max\left\{v_i\left(\sigma_{I\to a}^T, I\right) - v_i\left(\sigma^T, I\right), 0\right\} & T = 1 \\ \max\left\{R_i^{+,T-1}(I,a) + v_i\left(\sigma_{I\to a}^T, I\right) - v_i\left(\sigma^T, I\right), 0\right\} & T > 1 \end{cases}, \tag{11}$$

and strategies updates are weighted linearly in time i.e.

$$s_I[a] \leftarrow s_I[a] + \pi_{-i}[I]\sigma[a][I]w^T \tag{12}$$

where is the linear weighting term such that the weight of the update at any iteration T is given by $w^T = \max\{T-d, 0\}$, d being some iteration threshold after which we begin weighting linearly. This weighting scheme, explored more in discounted CFR [3], places higher importance on more recently sampled regrets and discounts the effects of regrets sampled in the past, yielding a tighter convergence bound.

While CFR+ is shown to yield considerable convergence rate improvements with vanilla CFR, this empirically is shown not to hold in outcome sampling [19]. However, when using variance reduction in conjunction with CFR+, namely VR-MCCFR+, convergence rate is improved relative to VR-MCCFR, thus we use VR-MCCFR+ for the dynamic orbit change game.

To assess the validity of a given strategy, we use the exploitability metric. Exploitability for a player is defined as the difference between the expected utility yielded by the current strategy and the utility that could be gained with a best response. Because with a Nash equilibrium strategy there is no incentive to deviate, the Nash equilibrium strategy will yield an exploitability of 0. Therefore, we can quantify the distance to a Nash equilibrium by exploitability.

The calculation of the best response for assessing exploitability reduces the problem from an imperfect information game to a partially observable Markov decision process (POMDP). While for small games the solution to this POMDP can be found with an exhaustive tree search, the orbit change game is too large to make this exhaustive search tractible. Therefore, we resort to the approximate POMDP solution method: partially observable Monte Carlo planning (POMCP) [22]. The tree constructed by the POMCP planner requires approximation of belief node values which get more accurate the more samples are taken. So, we can approximate the exploitability of the orbit change problem by the best response root belief value computed by the POMCP planner.

4. RESULTS

4.1 Mode Change Game Strategies

We used counterfactual regret minimization to solve the mode change game. Figure 1b shows the Nash equilibrium strategies of the both the ground station (in blue) and the satellite (in red) alongside a cardioid reward function (in green) with $N = 20$ orbit time segments, and a ground station budget of 7 scans. The action space for each player is

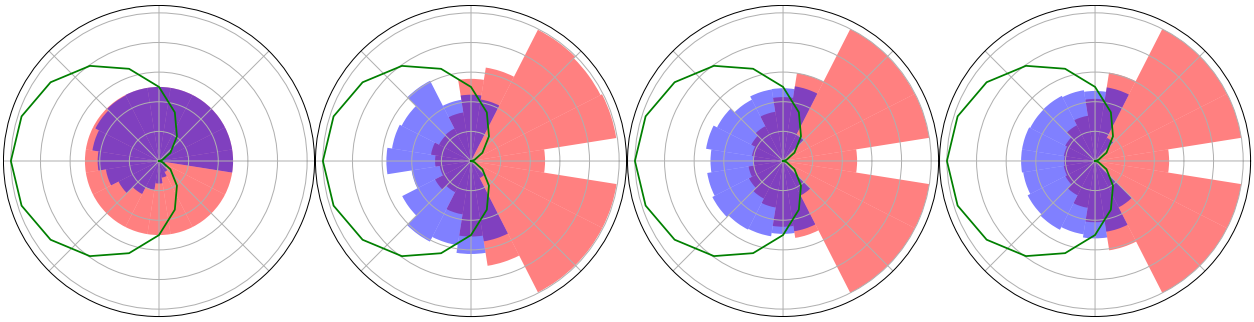


Fig. 2: Progression of polar strategy profile over training process. Provided strategies are snapshots at 0, 10, 10^3, and 10^4 ESCFR training iterations.

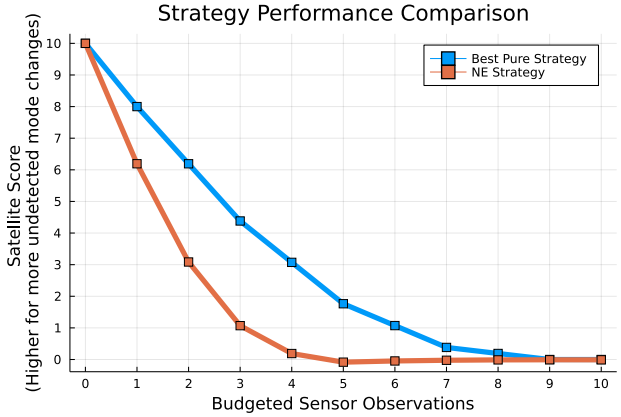


Fig. 3: Mode change game budget study

binary i.e. the ground station can either choose to scan ($a_g^{(s)}$) or wait ($a_g^{(w)}$) and similarly the satellite can choose to undergo a mode change ($a_s^{(c)}$) or wait ($a_s^{(w)}$). The red bars in the plot indicate the probability with which the satellite chooses to undergo a mode change at some point in its orbit, $P(a_s^{(c)}|t)$. Similarly, the blue bars in the polar bar plot indicate the probability with which the ground station will scan for some point in the satellite's orbit, $P(a_g^{(s)}|t)$[1].

Counterfactual regret methods are able to find a strategy profile that effectively distributes the ground station's budget probabilistically in a manner that "protects" higher-reward regions from the satellite. Consequently, we see that the satellite is relegated to mode changes in significantly lower reward regions on the right half of the polar plot.

Figure 2 shows the convergent nature of the ESCFR training process. Both strategies are initialized as uniformly random as shown on the far left. This uniform initialization causes the ground station to expend a majority of its budget in the beginning of the game leading to minimal scanning coverage over the later sectors. Over just 10 ESCFR iterations, ground station scanning coverage is shifted over to higher value sectors. By 10^3 training iterations a majority of high value sectors are covered by 50% scan probability. Finally, the results of 10^4 training iterations shows marginal improvement from 10^3, indicating reasonable convergence.

4.2 Scan Budget Analysis

We have demonstrated that, given some budget allotment, CFR can find a Nash equilibrium solution that maximizes expected utility under the worst-case opponent response. At a higher level, these game theoretic solutions can also

[1]Strictly speaking, the ground station's strategy is not only dependent on time t, but also the ground station's remaining budget b. Therefore, the true ground station strategy is expressed as $P(a_g|t,b)$, but for display, we marginalize out the budget condition, yielding the marginal scanning probability pictured in figure 1b

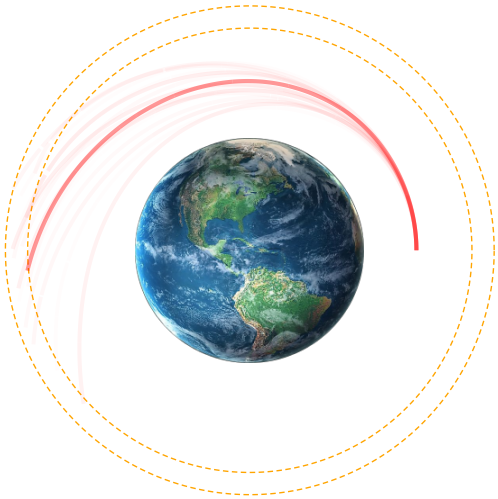


Fig. 4: Orbital trajectories in a satellite strategy calculated using VR-MCCFR for the Orbit Change Game. Opacity indicates the likelihood of the trajectory in the strategy.

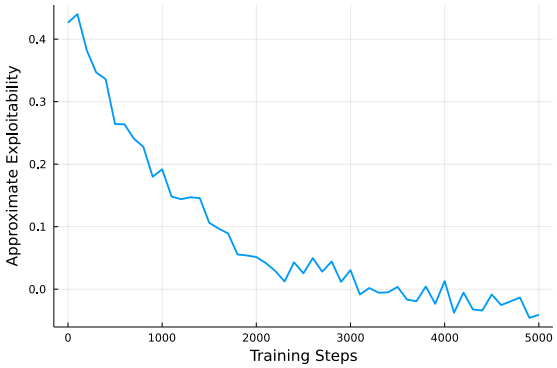


Fig. 5: Approximate exploitability of the orbit change game solution for 5000 training iterations.

answer a system design problem problem: What budget is required to achieve some desired worst-case expected utility, as demonstrated in Fig. 3. With an allotment of one scan to the ground station, the satellite naturally has almost free reign to undergo a mode change anywhere, likely with impunity. However, with a budget allotment that is half the total simulation time, the ground station is able to converge to a strategy that scans all sectors 50% of the time. Consequently, if the satellite were to undergo a mode change in a sector, the expected utility thereof would be 0, as it would be caught half the time. Ultimately, this forces an optimal strategy of the satellite at this budget to be to do nothing at all. In other words, despite being an active satellite, an optimal strategy for this satellite is to behave as inactive debris.

4.3 Orbit Change Game Strategies

Figure 4 demonstrates a snapshot of the satellite's strategy for which the satellite starts at a circular orbit at an altitude of 1.5 Earth radii and has a goal altitude of 2.1 Earth radii. We find that the satellite explores a set of feasible trajectories and then converges to an unexploitable mixture over these feasible trajectories.

In practice we find that VR-MCCFR+ works reasonably well to cover a majority of the game tree closer to the root while keeping the computation cost per training iteration low relative to external sampling and vanilla CFR. However, given the sparsity of outcome sampling, nodes in the game tree at a greater depth in the tree are either never visited or visited very rarely.

Nonetheless, Fig. 5 shows that the approximate exploitability of the calculated strategy decreases. While negative exploitability is not strictly possible, the negative approximate exploitability demonstrated in Fig. 5 simply shows that the POMCP planner was not able to generate a best response policy that yielded higher expected utility than that provided by VR-MCCFR+ for the allotted planning time of 5 seconds.

5. CONCLUSION

This paper has demonstrated game theory's advantages for SDA sensor tasking with limited observability. First, Monte Carlo Counterfactual regret methods are applied to a simple mode change game, yielding a stochastic strategy profile for a ground station that effectively disincentivizes a satellite from operating in high-value orbit sectors. Moreover,

we show that game-theoretic tasking significantly decreases the sensor time budget that must be allocated to fulfill this mission. Finally, we apply variance reduced outcome sampling CFR to a significantly larger orbit change game where satellite dynamics are governed by Keplerian dynamics. While external sampling works well for the mode change game, the relatively large size of the orbit change game requires the use outcome sampling with variance reduction. This method is shown to converge to a mixture over feasible goal-reaching trajectories for the satellite.

While the CFR methods used in this paper work well for the mode change and orbit change games, they fall short when applied to games with larger horizons or action spaces. To remedy this limitation, we plan to consider Policy Space Response Oracle [11] (PSRO) methods or DeepNash [17]. PSRO methods approximate solutions to extensive form games by mixing over a set of pure strategies and the empirically determined expected payoffs of these joint strategies. The set of pure strategies is expanded by adding the best response to the previous iteration's mixture. Because a best response strategy is a partially observable Markov decision process (POMDP), recent advances in POMDP solution methods can be leveraged to quickly calculate these best responses. On the other hand, DeepNash modifies the underlying game dynamics to quickly converge to an approximate Nash equilibrium without cycling around it, as is common with regret-based solution methods. This, coupled with neural network function approximation allows this method to extend to very large board games such as Stratego, without the need for game-specific heuristics.

6. ACKNOWLEDGEMENTS

The authors would like to thank L3 Harris Technologies for providing funding for this research and Derek Kingrey, Dustin Platter, and Galen Nickey for technical advice. This article contains the opinions and conclusions of its authors and not L3 Harris Technologies or its employees.

REFERENCES

[1] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Communications of the ACM*, 60(11):81–88, 2017.

[2] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. *CoRR*, abs/1811.00164, 2018.

[3] Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. *CoRR*, abs/1809.04040, 2018.

[4] Noam Brown, Tuomas Sandholm, and Strategic Machine. Libratus: The superhuman ai for no-limit poker. In *IJCAI*, pages 5226–5228, 2017.

[5] Neil Burch, Matej Moravcik, and Martin Schmid. Revisiting CFR+ and alternating updates. *CoRR*, abs/1810.11542, 2018.

[6] Samuel Fedeler, Marcus Holzinger, and William Whitacre. Sensor tasking in the cislunar regime using Monte Carlo tree search. *Advances in Space Research*, 70(3):792–811, 2022.

[7] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[8] Marcus Holzinger. Challenges and potential in space domain awareness. *Journal of Guidance, Control, and Dynamics*, 41(1):15–18, jan 2018.

[9] Andris D Jaunzemis, Marcus J Holzinger, and Moriba K Jah. Evidence-based sensor tasking for space domain awareness. In *Advanced Maui Optical and Space Surveillance Technologies Conference*, page 33. Maui Economic Development Board Maui, HI, 2016.

[10] Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte carlo sampling for regret minimization in extensive games. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[11] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[12] Richard Linares and Roberto Furfaro. Dynamic sensor tasking for space situational awareness via reinforcement learning. In *Advanced Maui Optical and Space Surveillance Tech. Conf.(AMOS)*, 2016.
[13] Richard Linares and Roberto Furfaro. An autonomous sensor tasking approach for large scale space object cataloging. In *Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS)*, pages 1–17, 2017.
[14] Stephen McAleer, Gabriele Farina, Marc Lanctot, and Tuomas Sandholm. Escher: Eschewing importance sampling in games by computing a history value function to estimate regret, 2022.
[15] Matej Moravčik, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.
[16] John F. Nash. Equilibrium points in ¡i¿n¡/i¿-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.
[17] Julien Perolat, Bart de Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of Stratego with model-free multiagent reinforcement learning, 2022.
[18] Martin Schmid. Search in imperfect information games. *CoRR*, abs/2111.05884, 2021.
[19] Martin Schmid, Neil Burch, Marc Lanctot, Matej Moravcik, Rudolf Kadlec, and Michael Bowling. Variance reduction in monte carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. *CoRR*, abs/1809.03057, 2018.
[20] Luke Schoenwetter. *Game Theory Applications in Astrodynamics and Space Domain Awareness*. PhD thesis, The University of Alabama, 2021. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-12-18.
[21] Luke Schoenwetter, Rohan Sood, and Brent Barbee. Optimal intercept of evasive spacecraft. 08 2020.
[22] David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems*, pages 2164–2172, 2010.
[23] Eric Steinberger, Adam Lerer, and Noam Brown. DREAM: deep regret minimization with advantage baselines and model-free learning. *CoRR*, abs/2006.10410, 2020.
[24] Zachary Sunberg, Suman Chakravorty, and Richard Erwin. Information space sensor tasking for space situational awareness. In *American Control Conference (ACC)*, pages 79–84, June 2014.
[25] Zachary Sunberg, Suman Chakravorty, and Richard Scott Erwin. Information space receding horizon control for multisensor tasking problems. *IEEE Transactions on Cybernetics*, 46(6):1325–1336, 2016.
[26] Oskari Tammelin. Solving large imperfect information games using CFR+. *CoRR*, abs/1407.5042, 2014.
[27] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20:1729–1736, 2007.