AIAA SciTech Forum
6-10 January 2020, Orlando, FL
AIAA Scitech 2020 Forum

10.2514/6.2020-1823

Approximate Optimal Orbit Transfer of Non-cooperative Debris

Max L. Greene*, Camilo Riano-Rios,†
Riccardo Bevilacqua‡, Norman G. Fitz-Coy,§
Warren E. Dixon¶
University of Florida, Gainesville, Florida, 32611

Motivated by the desire to use Coulomb forces as a non-contact means for one satellite to exert forces on another satellite for applications such as debris removal, a two-player non-cooperative zero-sum-game is formulated. Despite uncertainty in the dynamics of a disabled satellite (DS), including uncertainty in the interaction forces with a service satellite (SS), an approximately optimal real-time feedback policy (i.e., approximate dynamic programming (ADP)) is derived by using reinforcement learning and a Bellman error (BE) extrapolation method to identify the unknown value function provided a mild sufficient excitation condition is satisfied. In comparison to the typical pursuit-evade mini-max game, the SS is also tasked with regulating the orbital state of the DS to a desired state. A Lyapunov-based analysis is used to conclude that the SS approximately optimally intercepts the DS and regulates it to a neighborhood of the desired orbital state (i.e., uniformly ultimately bounded steady-state errors) while maintaining a desired relative offset distance. Initial simulations demonstrate the feasibility and performance of the developed controller using a potential field to model the uncertain interaction forces. Additional efforts will focus on a higher fidelity model of the Coulomb forces.

I. Introduction

Since the launch of Sputnik in 1957, there have been more than 5,000 space launches. Currently, there are more than 23,000 tracked objects in Earth orbits, of which only 1,200 can be considered functional spacecraft. Furthermore, based on models from the National Aeronautics and Space Administration and European Space Agency, it is estimated that over 750,000 objects larger than 1 cm and 100 million objects larger than 1 mm reside in Earth orbits, all of which are considered space debris.[1,2]

The international space community has established several guidelines and policies to address the growth in orbital debris, which is a major concern due to the risk it poses to payloads and onboard instruments/electronics. For example, in 2007 the United Nations adopted the "25-year rule" space debris mitigation guidelines as a mechanism to address this concern.[3] However, some recent studies have shown that the debris population in low Earth orbit (LEO) may have reached the point where the debris population will continue to grow in the next 200 years, even without any future launches.[4,5] While these studies focused on the LEO region, the medium Earth orbit, and the geostationary Earth orbit (GEO) regions also have significant space debris concerns, particularly due to the mega-constellations that are being proposed. In response to these concerns, active debris removal (ADR) has been proposed as a viable approach to mitigate the growth of space debris.[6]

The Inter-Agency Space Debris Coordination Committee has established two approaches for ADR: deorbiting to reentry for low Earth orbits or re-orbiting to graveyard regions for higher altitude orbits.[7] Conceptually, ADR can be broadly decomposed into methods which involve (i) passive/active devices onboard

*Graduate Research Assistant, Department of Mechanical and Aerospace Engineering; maxgreene12@ufl.edu.
†Graduate Research Assistant, Department of Mechanical and Aerospace Engineering; crianorios@ufl.edu.
‡Associate Professor, Department of Mechanical and Aerospace Engineering; bevilr@ufl.edu.
§Associate Professor, Department of Mechanical and Aerospace Engineering; ncf@ufl.edu.
¶Professor, Department of Mechanical and Aerospace Engineering; wdixon@ufl.edu.

the spacecraft for post-mission disposal, (ii) sublimation of the debris, or (iii) "capture" and maneuvering of the debris.[8,9] Methodology (i) involves devices integrated into the spacecraft and thus are applicable to future missions; methodology (ii) typically involves high powered ground-based or space-based lasers and pose significant technical and legal challenges. Methodology (iii) is by far the most practical ADR for existing debris and have received significant interest from the community. Methodology (iii) can be further decomposed into contact or non-contact capture approaches, where the prevailing ideas of the non-contact approach is the "electrostatic tractor". For example, previous results[10–17] have examined the use of Coulomb forces (e.g., using EP) as a non-contact means of one satellite exerting forces on another satellite for missions such as debris removal, formation management, etc.

This paper examines the electrostatic tractor problem via a two-player non-cooperative zero-sum-game despite uncertainty in the disabled satellite (DS), including the uncertain non-contact interaction force function. Motivated by the desire to develop a feedback policy, an approximate dynamic programming (ADP) method inspired by our general result[18] is pursued for an infinite horizon control affine mini-max game, where reinforcement learning methods are used to approximate the unknown value function (i.e., the solution to the Hamilton-Jacobi-Isaacs (HJI) equation).[19,20] To address computational concerns, a state following (StaF)[21,22] neural network architecture is used. Moreover, the Bellman error (BE) resulting from the use of actor-critic estimates of the value function is evaluated through a simulation of experience approach[23] to yield simultaneous exploration and exploitation. This approach yields a finite time sufficient excitation condition (as opposed to a persistent excitation (PE)) that can be checked online from an eigenvalue condition. Unlike typical pursuit-evasion and reach-avoid games,[23–41] the electrostatic tractor problem involves an additional step of regulation to a new orbital state after target interception of the non-cooperative player. A Lyapunov-like stability analysis is performed to conclude that the optimal policy is approximated and the orbital state is changed to a desired state within some residual error (i.e., uniformly ultimately bounded (UUB) convergence). Initial simulations are provided using a potential field to represent distance based Coulomb forces between the service satellite (SS) and DS. Additional efforts will focus on also compensating for uncertain SS dynamics, including more realistic electrostatic forces in the simulation, and generalizing the development to consider the ability of electrostatic forces to pull and push the DS (e.g., the current paper relies on the SS to move in a manner that accounts for one-directional Coulomb forces).

II. Problem Formulation

The objective is for a SS to approach and impart forces on a DS to change the orbital state of the DS, denoted by $z : \mathbb{R}_{t\geq t_0} \to \mathbb{R}^n$, to a desired orbital state, denoted by$z_g \in \mathbb{R}^n$, while minimizing the cost to perform such an action using a feedback policy, despite uncertainty in the DS dynamics and the interaction between the SS and DS. The error terms $e_1 \in \mathbb{R}^n$ and $e_2 \in \mathbb{R}^n$, defined as

$$e_1 \triangleq z(t) - (\eta(t) - r_d) \tag{1}$$

$$e_2 \triangleq ((\eta(t) - r_d) - z_g) - k_1 (z(t) - z_g). \tag{2}$$

The error systems described in (1) and (2) quantify the objective of the SS to intercept the DS at a constant standoff distance[a], denoted by $r_d \in \mathbb{R}^n$, and the SS's objective to regulate the DS's orbital state, respectively, where $\eta : \mathbb{R}_{t\geq t_0} \to \mathbb{R}^n$ denotes the orbit state of the SS, $k_1 \in \mathbb{R}_{>1}$ is a positive constant, and $t_0 \in \mathbb{R}_{\geq 0}$ is the initial time. To facilitate the subsequent development, after some algebraic manipulation, the DS and SS orbital states can be expressed in terms of the errors in (1) and (2) as

$$z(t) = \frac{1}{(1-k_1)}(e_1 + e_2) + z_g, \tag{3}$$

$$\eta(t) = \frac{1}{(1-k_1)}(k_1 e_1 + e_2) + r_d + z_g. \tag{4}$$

$$\dot{z}(t) = f(z(t), \eta(t), t), \tag{5}$$

[a]The standoff distance is a user defined term that that is sufficiently small enough to allow the SS to exert orbital correction forces on the DS while avoiding collision.

$$\dot{\eta}(t) = h(\eta(t)) + g(z(t), \eta(t)) u(t), \tag{6}$$

respectively, where $f : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_{t \geq t_0} \to \mathbb{R}^n$ denotes the unknown drift dynamics which are locally Lipschitz in the states and continuous in t, $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ denotes the unknown locally Lipschitz continuous drift dynamics, $g : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ is the known control effectiveness matrix, and $u : \mathbb{R}_{\geq t_0} \to \mathbb{R}^m$ is the control input for the SS. The structure of (5) represents the interaction forces imparted by SS that depend on the state of the DS and an additional exogenous dynamic (e.g., time-varying forces exerted on the DS from an EP system on the SS). For simplicity, and as often done in min-max differential game literature,[19,20] the dynamics in (5) are expressed as an equivalent exogenous disturbance $d : \mathbb{R}_{\geq t_0} \to \mathbb{R}^n$ given by

$$\dot{z}(t) = d(t) \triangleq f(z(t), \eta(t), t). \tag{7}$$

The dynamics in (6) and (7) can be combined into a compact form that facilitates the subsequent development as

$$\dot{x}(t) = F(x(t)) + G(x(t)) u(t) + K d(t), \tag{8}$$

where $x(t) \triangleq \left[e_1^T(t), e_2^T(t)\right]^T \in \mathbb{R}^{2n}$ is a concatenated state vector. $F : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$, $G : \mathbb{R}^{2n} \to \mathbb{R}^{2n \times m}$, $K \in \mathbb{R}^{2n \times n}$ are defined as

$$F(x(t)) \triangleq \begin{bmatrix} -h(s_1(x(t)), s_2(x(t))) \\ h(s_1(x(t)), s_2(x(t))) \end{bmatrix}, \tag{9}$$

$G^T(x(t)) \triangleq \begin{bmatrix} -g(s_1(x(t)), s_2(x(t)))^T, & g(s_1(x(t)), s_2(x(t)))^T \end{bmatrix}^T$, and $K^T \triangleq \begin{bmatrix} I_n & , -k_1 I_n \end{bmatrix}^T$, where s_1, $s_2 : \mathbb{R}^{2n} \to \mathbb{R}^n$ are auxiliary terms defined as $s_1(x(t)) \triangleq z(t)$ and $s_2(x(t)) \triangleq \eta(t)$. The dynamics of $F(x(t))$ are assumed to be locally Lipschitz continuous, and, hence, $\|F(x)\| \leq L \|x\|$ where $L \in \mathbb{R}_{\geq 0}$ is the Lipschitz constant.

III. Approximate Optimal Control Development

To quantify the trade-off between the goal of changing the orbital state and the control effort required to achieve the goal, a cost function is formulated as a two-player zero-sum game with the SS controller being the minimizing player and the disturbance dynamics of the DS being the maximizing player. Specifically, the infinite horizon quadratic cost function is defined as

$$J(x, u, d, t_0) \triangleq \int_{t_0}^{\infty} r(x(\tau), u(\tau), d(\tau)) d\tau, \tag{10}$$

which is under the constraint of the dynamics in (8), where the instantaneous cost function $r : \mathbb{R}^{2n} \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is defined as $r(x(t), u(t), d(t)) \triangleq Q(x(t)) + u^T(t) R u(t) - \gamma^2 d^T(t) d(t)$, where $Q : \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$ is a user-defined[b] positive definite weight matrix of the states (i.e., $\underline{q} \|x\|^2 \leq Q(x) \leq \overline{q} \|x\|^2$ where $\underline{q}, \overline{q} \in \mathbb{R}_{>0}$ are known positive constants), $R \in \mathbb{R}^{m \times m}$ is a positive definite symmetric weight matrix of the control input, and $\gamma \in \mathbb{R}_{>0}$ is a constant gain. Motivated to develop a feedback policy (i.e., rather than a numerical solution that is executed in open-loop over a finite horizon), a dynamic programming approach is pursued. However, given the uncertain dynamics of the DS, and the fact that the optimal value function $V^* : \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$, defined as

$$V^*(x(t)) \triangleq \min_{u(\tau)} \max_{d(\tau)} \int_t^{\infty} r(x(\tau), u(\tau), d(\tau)) d\tau, \tag{11}$$

is unknown, an ADP solution is sought. Specifically, the optimal value function is characterized by the HJI equation [31,42–45]

$$\begin{aligned} 0 = {} & r(x(t), u^*(x(t)), d^*(t)) + \nabla V^*(x(t)) F(x(t)) \\ & + \nabla V^*(x(t)) (G(x(t)) u^*(x(t)) + K d^*(x(t))), \end{aligned} \tag{12}$$

[b] For example, let $Q(x(t)) \triangleq x^T(t) Q_c x(t)$ with the positive definite symmetric constant matrix $Q_c \in \mathbb{R}^{2n \times 2n}$.

for all $t \in \mathbb{R}_{\geq t_0}$ with $V^*(0) = 0$, where ∇V is defined as $\nabla V \triangleq \frac{\partial V(x)}{\partial x}$.[c] If a solution to (12) exists in which $V^*(x) \geq 0$, then the optimal SS control policy and worst-case DS disturbing policy can be determined by solving

$$0 = \min_u \max_d \left(r\left(x\left(t\right), u^*\left(x\left(t\right)\right), d^*\left(t\right)\right) + \nabla V^*\left(x\left(t\right)\right)\left(F\left(x\left(t\right)\right) + G\left(x\left(t\right)\right)u^*\left(x\left(t\right)\right) + K\left(x\left(t\right)\right)d^*\left(x\left(t\right)\right)\right)\right),$$

where the optimal SS policy, $u^* : \mathbb{R}^{2n} \to \mathbb{R}^m$, is

$$u^* = -\frac{1}{2}R^{-1}G^T\left(x\right)\left(\nabla V^*\left(x\left(t\right)\right)\right)^T, \tag{13}$$

and the worst-case DS policy, $d^* : \mathbb{R}^{2n} \to \mathbb{R}^n$, is

$$d^*\left(x\right) = \frac{1}{2\gamma^2}K^T\nabla V^*\left(x\left(t\right)\right)^T. \tag{14}$$

Motivated by computational concerns (i.e., the Bellman curse of dimensionality) and previous results,[22] computationally efficient state-following (StaF) kernels are used to approximate the unknown value function. To facilitate the StaF kernel development, let $\overline{B_r\left(x\right)}$ be the closure of an open ball of radius $r \in \mathbb{R}_{>0}$ centered at x, with $y \in \overline{B_r\left(x\right)}$ and $c\left(x\right) \in \overline{B_r\left(x\right)}$, where $c : \chi \to \chi^P$ are state-following centers around $x \in \chi$ for the compact domain $\chi \subset \mathbb{R}^{2n}$. The notation $\overline{\|(\cdot)\|}$ is defined as $\overline{\|(\cdot)\|} \triangleq \sup_{\xi \in B_\zeta} \|(\cdot)\|$. Specifically, the optimal value function, optimal SS policy, and worst-case DS policy can be expressed as

$$V^*\left(x\right) = W\left(x\right)^T\sigma\left(c\left(x\right)\right) + \epsilon_W\left(x\right), \tag{15}$$

$$u^*\left(x\right) = -\frac{R^{-1}G^T\left(x\right)}{2}\left(\nabla\sigma\left(c\left(x\right)\right)^T W\left(x\right) + \epsilon_W\left(x\right)^T\right), \tag{16}$$

$$d^*\left(x\right) = \frac{K^T}{2\gamma^2}\left(\nabla\sigma\left(c\left(x\right)\right)^T W\left(x\right) + \epsilon_W^T\left(x\right)\right), \tag{17}$$

respectively, where $W : \chi \to \mathbb{R}^P$ is the vector of continuously differentiable ideal StaF weights, $\sigma : \chi \times \chi \to \mathbb{R}^P$ is a bounded vector of continuously differentiable nonlinear kernels, $\epsilon_v : \chi \times \chi \to \mathbb{R}$ is a continuously differentiable function approximation error, and $\epsilon_W\left(x, y\right) \triangleq \sigma\left(y, c\left(x\right)\right)^T\nabla W\left(x\right) + \nabla\epsilon_v\left(x, y\right)$. Based on the structure of (15), $\|\nabla V^*\left(x\right)\| \leq \alpha\|x\|\ \forall x \in \mathbb{R}^n$, where $\alpha \in \mathbb{R}_{>0}$ is a constant [d]. By invoking the Stone-Weierstrass approximation theorem,[47] approximations for (15)-(17), denoted by $\hat{V} : \mathbb{R}^{2n} \times \mathbb{R}^P \to \mathbb{R}$, $\hat{u} : \mathbb{R}^{2n} \times \mathbb{R}^P \to \mathbb{R}^m$, and $\hat{d} : \mathbb{R}^{2n} \times \mathbb{R}^P \to \mathbb{R}^n$, respectively, are

$$\hat{V}\left(x, \hat{W}_c\right) = \hat{W}_c^T\sigma\left(c\left(x\right)\right), \tag{18}$$

$$\hat{u}\left(x, \hat{W}_a\right) = -\frac{R^{-1}G^T\left(x\right)}{2}\nabla\sigma\left(c\left(x\right)\right)^T\hat{W}_a, \tag{19}$$

$$\hat{d}\left(x, \hat{W}_d\right) = \frac{K^T}{2\gamma^2}\nabla\sigma\left(c\left(x\right)\right)^T\hat{W}_d, \tag{20}$$

where $\hat{W}_c, \hat{W}_a, \hat{W}_d \in \mathbb{R}^P$ denote the critic, actor, and disturbance weight estimates, respectively. After replacing (15)-(17) with (18)-(20), the resulting Bellman error (BE) (i.e., a measure of optimality[23]) is

$$\delta_t\left(x, \hat{W}_c, \hat{W}_a, \hat{W}_d\right) = r\left(x, \hat{u}\left(x, \hat{W}_a\right), \hat{d}\left(x, \hat{W}_d\right)\right) \tag{21}$$

$$+ \nabla\hat{V}\left(x, \hat{W}_c\right)\left(F\left(x\right) + G\left(x\right)\hat{u}\left(x, \hat{W}_a\right) + K\hat{d}\left(x, \hat{W}_d\right)\right), \tag{22}$$

[c] The HJI in (12) requires knowledge of the DS and SS drift dynamics. Future work will utilize system identification to characterize the drift dynamics using the universal function approximation property.[46]

[d] Previous results[29] assume that $\nabla V^*\left(x\right)$ is bounded by a constant as $\|\nabla V^*\left(x\right)\| \leq \alpha$. This assumption is generalized in this paper due to the use of the additional state dependent terms from the StaF method.

where $\delta_t : \mathbb{R}^{2n} \times \mathbb{R}^P \times \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}$. The subscript t is used to indicate that (22) is the instantaneous form of the BE. Specifically, it is the on-policy BE evaluated at each time instance with the current weight parameter estimates $\hat{W}_c(t), \hat{W}_a(t)$, and $\hat{W}_d(t)$. If only the instantaneous BE were used to update the weight estimates, then the traditional integral-based persistence of excitation (PE) requirement[48] would be required to identify the value function. Motivated by the desire to relax the integral PE condition, a learning by simulation of experience strategy[23,49,50] is used for off-policy evaluation of the BE in the region of $\overline{B_r(x)}$, yielding virtual excitation. The extrapolated BE, $\delta_{ti}(t)$, includes the subscript i to indicate the i^{th} extrapolation point, and is otherwise defined as in (22) with $\hat{W}_{ci}(t), \hat{W}_{ai}(t)$, and $\hat{W}_{di}(t)$. The instantaneous and extrapolated BE forms provide (on-policy and off-policy) measures of optimality that are used to adjust the actor and critic estimates in (18)-(20) through the following update laws

$$\dot{\hat{W}}_c(t) = -\Gamma_c(t)\frac{k_{c1}}{N+1}\frac{\omega(t)}{\rho^2(t)}\delta_t(t) - \Gamma_c(t)\frac{k_{c2}}{N+1}\sum_{i=1}^{N}\frac{\omega_i(t)}{\rho_i^2(t)}\delta_{ti}(t), \tag{23}$$

$$\dot{\hat{W}}_a(t) = \text{proj}\left\{-K_a k_{a1}\left(\hat{W}_a(t) - \hat{W}_c(t)\right)\right\}, \tag{24}$$

$$\dot{\hat{W}}_d(t) = \text{proj}\left\{-K_d k_{d1}\left(\hat{W}_d(t) - \hat{W}_c(t)\right)\right\}, \tag{25}$$

where $\omega(t) \triangleq \nabla\sigma(x(t), c(x(t)))\left(F(x(t)) + G(x(t))\hat{u}\left(x(t), \hat{W}_a(t)\right) + K\hat{d}\left(x(t), \hat{W}_d(t)\right)\right)$, $\Gamma_c(t) \in \mathbb{R}^{P\times P}$ is a positive definite least squares gain matrix updated by[e]

$$\dot{\Gamma}_c(t) = \beta_c\Gamma_c(t) - \Gamma_c(t)\frac{k_{c1}}{N+1}\frac{\omega(t)\omega^T(t)}{\rho^2(t)}\Gamma_c(t) - \Gamma_c(t)\frac{k_{c2}}{N+1}\sum_{i=1}^{N}\frac{\omega_i(t)\omega_i^T(t)}{\rho_i^2(t)}\Gamma_c(t), \tag{26}$$

where $\rho(t) \triangleq \sqrt{1+\gamma_1\omega^T(t)\omega(t)}$ and $\rho_i(t) \triangleq \sqrt{1+\gamma_1\omega_i^T(t)\omega_i(t)}$, $k_{c1}, k_{c2}, \gamma_1, \beta_c, k_{a1}, k_{d1} \in \mathbb{R}_{>0}$ are learning gains, $K_a, K_d \in \mathbb{R}^{P\times P}$ are positive definite symmetric gain matrices, and $N \in \mathbb{R}$ is the number of StaF kernels. The advantage of the learning by simulation of experience strategy and the resulting updates laws in (23)-(26) is that the traditional integral PE condition (see (29)) can be relaxed to the broader set of sufficient excitation conditions (see (27)-(29)) in the following assumption, which can be validated on-line.

Assumption 1. [23,50] *There exists a strictly positive $T \in \mathbb{R}_{>0}$ and at least one strictly positive constant $\underline{c}_1, \underline{c}_2 \in \mathbb{R}_{\geq 0}$, or $\underline{c}_3 \in \mathbb{R}_{\geq 0}$ such that*

$$\underline{c}_1 I_L \leq \inf_{t\in\mathbb{R}_{\geq t_0}} \frac{1}{N+1}\sum_{i=1}^{N}\frac{\omega_i(t)\omega_i^T(t)}{\rho_i^2(t)}, \tag{27}$$

$$\underline{c}_2 I_L \leq \int_t^{t+T}\left(\frac{1}{N+1}\sum_{i=1}^{N}\frac{\omega_i(\tau)\omega_i^T(\tau)}{\rho_i^2(\tau)}\right)d\tau,\ \forall t \in \mathbb{R}_{\geq t_0}, \tag{28}$$

$$\underline{c}_3 I_L \leq \int_t^{t+T}\left(\frac{1}{N+1}\frac{\omega(\tau)\omega^T(\tau)}{\rho^2(\tau)}\right)d\tau,\ \forall t \in \mathbb{R}_{\geq t_0}. \tag{29}$$

IV. Stability Analysis

A Lyapunov-based stability analysis is provided in this section to examine the convergence of a stacked vector of the closed-loop orbital transfer and function approximation errors defined as $Z_L \triangleq \left[x^T, \tilde{W}_c^T, \tilde{W}_a^T, \tilde{W}_d^T\right]^T \in \mathbb{R}^{2n+3P}$, where the function approximation error notation $\tilde{(\cdot)} \triangleq (\cdot) - \hat{(\cdot)}$. To facilitate the subsequent analysis,

[e] The gain matrix $\Gamma_c(t)$ is positive definate ($\underline{\Gamma}_c I_P \leq \Gamma_c(t) \leq \overline{\Gamma}_c I_P$, where $\underline{\Gamma}_c, \overline{\Gamma}_c \in \mathbb{R}_{>0}$ are constant bounds) provided $\lambda_{\min}\{\Gamma_c(t_0)\} > 0$.[51]

the auxiliary terms $\kappa, \iota \in \mathbb{R}_{>0}$ are defined as $\kappa \triangleq \frac{1}{8}\min\left\{\underline{q}, k_{c2}\underline{c}, k_{a1}, k_{d1}\right\}$, and $\iota \triangleq \frac{\iota_x^2}{\underline{q}} + \frac{\iota_c^2}{k_{c2}} + \frac{\iota_a^2}{k_{a1}} + \frac{\iota_d^2}{k_{d1}}$. The constants $\underline{c}, \iota_x, \iota_c, \iota_a, \iota_d \in \mathbb{R}_{>0}$ are defined as $\underline{c} \triangleq \left(\frac{\beta_c}{2k_{c2}\overline{\Gamma}_c} + \frac{\underline{c}_1}{2}\right)$,

$$\iota_x \triangleq \frac{\alpha}{2}\left\|K_\gamma \nabla\sigma^T \hat{W}_d - \frac{1}{2}G_R \nabla\sigma^T \hat{W}_a\right\|,$$

$$\iota_c \triangleq \frac{1}{\underline{\Gamma}_c}\|\nabla W\|\left(L\|x\| + \left\|\frac{1}{2}K_\gamma \nabla\sigma^T \hat{W}_d - \frac{1}{2}G_R \nabla\sigma^T \hat{W}_a\right\|\right)$$
$$+ \frac{\max\{k_{c1}, k_{c2}\}}{2\sqrt{\gamma_c}}\left(\overline{\|\Delta\|} + \frac{1}{4}\overline{\left\|\tilde{W}_a\right\|}\,\overline{\|G_\sigma\|}\,\overline{\left\|\tilde{W}_a\right\|} + \frac{1}{4}\overline{\left\|\tilde{W}_d\right\|}\,\overline{\|K_\sigma\|}\,\overline{\left\|\tilde{W}_d\right\|}\right),$$

$$\iota_a \triangleq \frac{1}{\lambda_{\min}\{K_a\}}\|\nabla W\|\left(L\|x\| + \left\|\frac{1}{2}K_\gamma \nabla\sigma^T \hat{W}_d - \frac{1}{2}G_R \nabla\sigma^T \hat{W}_a\right\|\right),$$

and

$$\iota_d \triangleq \frac{1}{\lambda_{\min}\{K_d\}}\|\nabla W\|\left(L\|x\| + \left\|\frac{1}{2}K_\gamma \nabla\sigma^T \hat{W}_d - \frac{1}{2}G_R \nabla\sigma^T \hat{W}_a\right\|\right),$$

where $G_\sigma \triangleq \nabla\sigma G_R \nabla\sigma^T$, $K_\sigma \triangleq \nabla\sigma K_\gamma \nabla\sigma^T$, $G_R \triangleq GR^{-1}G^T$, $K_\gamma \triangleq \frac{1}{\gamma^2}KK^T$, and

$$\Delta \triangleq \epsilon_W\left(\frac{1}{4}GR^{-1}G^T - \frac{1}{4}\frac{1}{\gamma^2}KK^T\right)\epsilon_W^T - \epsilon_W F + \frac{1}{2}\epsilon_W GR^{-1}G^T\nabla\sigma^T W - \frac{1}{2}\epsilon_W\frac{1}{\gamma^2}KK^T\nabla\sigma^T W,$$

so that $\overline{\|\Delta\|}$ and $\overline{\|\Delta_i\|}$ decrease as $\overline{\|\nabla W\|}$, $\overline{\|\nabla\epsilon_v\|}$, and $\overline{\|\nabla\epsilon_{v,i}\|}$ decrease.

Theorem 1. *Provided Assumption 1 is satisfied and the following sufficient conditions hold*

$$\underline{q} \geq \frac{1}{2}\alpha^2\left\|\left(GR^{-1}G^T - \frac{1}{\gamma^2}KK^T\right)\right\|, \tag{30}$$

$$\underline{c} \geq \frac{k_{a1} + k_{d1}}{k_{c2}}, \tag{31}$$

$$\sqrt{\frac{\iota}{\kappa}} \leq \underline{v}_l^{-1}\left(\overline{v}_l\left(\zeta\right)\right), \tag{32}$$

then the feedback policies in (18)-(20) *and update laws* in *(23)-(26) are bounded, and the closed-loop error systems are uniformly ultimately bounded in the sense that [52, Theorem 4.18]*

$$\limsup_{t\to\infty}\|Z_L(t)\| \leq \underline{v}_l^{-1}\left(\overline{v}_l\left(\sqrt{\frac{\iota}{\kappa}}\right)\right). \tag{33}$$

Proof. Consider the Lyapunov function candidate $V_L : \mathbb{R}^{2n+3P} \times \mathbb{R}_{\geq t_0} \to \mathbb{R}_{\geq 0}$ defined as

$$V_L(Z_L, t) \triangleq V^*(x) + \frac{1}{2}\tilde{W}_c^T\Gamma_c^{-1}(t)\tilde{W}_c + \frac{1}{2}\tilde{W}_a^T K_a^{-1}\tilde{W}_a + \frac{1}{2}\tilde{W}_d^T K_d^{-1}\tilde{W}_d, \tag{34}$$

which can be bounded by class $\mathcal{K}$ functions $\underline{v}_l, \overline{v}_l : \mathbb{R} \to \mathbb{R}_{\geq 0}$ as

$$\underline{v}_l\left(\|Z_L\|\right) \leq V_L(Z_L, t) \leq \overline{v}_l\left(\|Z_L\|\right). \tag{35}$$

Taking the time-derivative of (34) along its trajectory results in

$$\dot{V}_L(Z_L,t) = \nabla V^* \left(F + G\hat{u} + K\hat{d}\right) + \tilde{W}_c^T \Gamma_c^{-1} \left(\dot{W} - \dot{\hat{W}}_c\right) \tag{36}$$
$$- \frac{1}{2}\tilde{W}_c^T \left(\Gamma_c^{-1}\dot{\Gamma}_c\Gamma_c^{-1}\right)\tilde{W}_c + \tilde{W}_a^T K_a^{-1}\left(\dot{W} - \dot{\hat{W}}_a\right) + \tilde{W}_d^T K_d^{-1}\left(\dot{W} - \dot{\hat{W}}_d\right)$$

Substituting (22)-(25), using Assumption 1, and performing some algebraic manipulation results in the following inequality

$$\dot{V}_L(Z_L,t) \leq -\kappa \|Z_L\|^2 - \left(\kappa \|Z_L\|^2 - \iota\right) \tag{37}$$

provided the sufficient conditions in (30)-(32) are satisfied. Based on (37) the result in (33) can be obtained. Furthermore, from (34) and (37), we can conclude that $\|Z_L\| \in \mathcal{L}_\infty$. Using this fact, the approximate controller can shown to be bounded. □

V. Simulation Results

An initial simulation with several simplifying assumptions is used to demonstrate the feasibility of the developed controller and estimation methods. Since Coulomb forces (e.g., using the Debye-Huckel approximation[53]) are a function of the distance between the two bodies, the dynamics of the DS were modeled as a disturbed potential field as

$$\dot{z}(t) = A\Phi_1(x(t)) + B\Phi_2(x(t)) + \omega_r(t), \tag{38}$$

where $A \triangleq \begin{bmatrix} 12 & -4.8 & 8 \\ 6 & 12 & 2 \\ -12 & 4 & 3.6 \end{bmatrix}$, $B \triangleq \begin{bmatrix} 0.6 & -6 & 4.5 \\ 7.2 & 0.3 & 3 \\ -4.5 & 2.7 & 7.5 \end{bmatrix}$, $\Phi_1(x(t)) \triangleq (z(t) - \eta(t))\exp\left(\frac{-|z(t)-\eta(t)|^2}{2}\right)$, $\Phi_2(x(t)) \triangleq (z(t) - z_g)\exp\left(\frac{-|z(t)-z_g|^2}{8}\right)$, and $\omega_r(t)$ is an exogenous disturbance selected from a uniform distribution of $U[0,5]1_{3\times 1}$ at each time step. The SS dynamics in (6) are selected as $h(z(t),\eta(t)) \triangleq \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$, and $g(\eta(t)) = I_3$. The orbital components of Hill's reference frame are ignored and the control input into the SS is the velocity of the spacecraft given by $u(t) \triangleq \begin{bmatrix} u_x & u_y & u_z \end{bmatrix}^T$,where u_x is the commanded x-component of the velocity of the SS, u_y is the commanded y-component of the velocity of the SS, and u_z is the commanded z-component of the velocity of the SS. The x-axis points from the center of the Earth to the origin of the system (which moves along a circular orbit), the y-axis is along the orbital track, and the z-axis completes a right-hand Cartesian coordinate system. While ignoring the orbital components of of Hill's reference frame is unrealistic, the assumption is made to bifurcate the controller and spacecraft dynamics to quantify the controller performance. Additional simulation efforts target spacecraft dynamics in the Hill reference frame and use the force generated by the spacecraft's thrusters as a control input. Any spin induced on either the SS or DS is assumed to be internally stabilized using reaction wheels or control moment gyroscopes. This simplifying assumption clearly does not hold for problems such as de-tumbling space debris without including the electrostatic torques induced on the spacecraft. However, for the purposes of this development, torques are ignored to simplify the problem and show a proof-of-concept that an ADP-based controller can be used to herd EP satellite systems.

For function approximation, the StaF basis is selected as $(x, c(x)) = [\sigma_1(x, c_1(x)), \ldots, \sigma_7(x, c_7(x))]^T$, where $\sigma_i(x(t), c_i(x(t))) = x^T(t)(x(t) + 0.05\nu(x(t))d_i)$, $\nu(x(t)) = \frac{x^T x}{(1+x^T x)}$, and d_i are the vertices of a 4-simplex. To perform BE extrapolation, a single trajectory $x_i(x(t),t)$ is selected at random from a uniform distribution over a $0.05\nu(x(t)) \times 0.05\nu(x(t))$ square centered at the current state $x(t)$. Three points were used in BE extrapolation.

Initial conditions, user defined cost parameters, and learning gains given in Table 1, the SS successfully regulates itself to an offset from the goal location, $z_g + r_d$ and regulates the DS to the goal location, z_g. The trajectories of the DS and SS are shown in Figure 1. The concatenated error states, $x(t)$, are shown to converge to the origin in Figure 2, and RMS error values of each state are shown in Table 2: different results

would result from different initial conditions of the SS and DS. The states of the SS and DS are shown in Figures 3 and 4, respectively. These figures indicate that the SS converges toward z_g in approximately five seconds, and tracks the DS over a smaller distance than it did in the initial five seconds. Figure 5 shows the error between the magnitude of r_d and $\|\eta(t) - z(t)\|$. Furthermore, the difference in states, $\eta(t) - z(t)$, is depicted in Figure 6, indicating the desired standoff distance is maintained. The control effort is plotted in Figure 7. After intercepting the DS, the SS has a nonzero control input due to the disturbance in (38) and the nonzero interaction force between the SS and DS. Figure 7 shows the DS disturbing policy that governs the disturbing input of the augmented system dynamics in (8).

Table 1. Simulation initial conditions and parameters.

Initial conditions at $t_0 = 0$
$z(0) = [-3, 2, 2]^T$, $\eta(0) = [5, 5, -5]^T$, $r_d = [-1, -0.75, 0.5]^T$, $z_g = [1, 0.75, -0.5]^T$, $\hat{W}_c(0) = 0.1 \times 1_{7\times1}$, $\hat{W}_a(0) = 100 \times 1_{7\times1}$, $\hat{W}_d(0) = 100 \times 1_{7\times1}$,$\Gamma_c(0) = 0.5I_7$.
Penalizing parameters
$r(x(t), u(t), d(t)) \triangleq x^T Q_x x + u^T(t) R u(t) - \gamma^2 d^T(t) d(t)$, $Q_x = \text{diag}\{100, 100, 100, 100, 100, 100\}$, $R = \text{diag}\{0.09, 0.09, 0.09\}$, $\gamma = 2$.
Gains and parameters for ADP update laws
$k_{c1} = 0.01$, $k_{c2} = 0.75$, $k_{a1} = 0.5$, $k_{a2} = 0.75$, $k_{d1} = 0.5$, $k_{d2} = 0.005$, $\beta = 0.01$, $\beta_c = 0.001$,$K_a = I_5$, $K_d = I_5$.

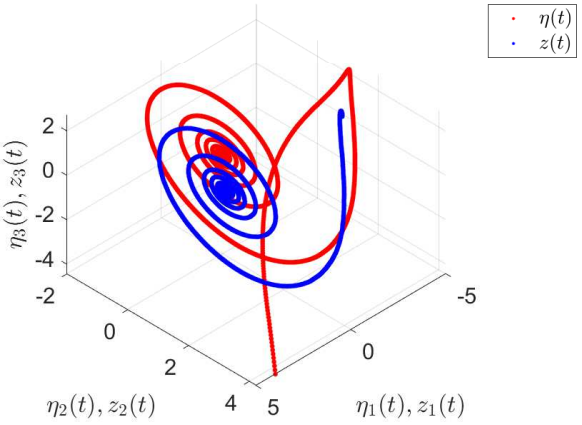

Figure 1. Trajectory of the SS and DS. The SS (red dot) intercepts and drives the DS (blue dot) to the desired state.

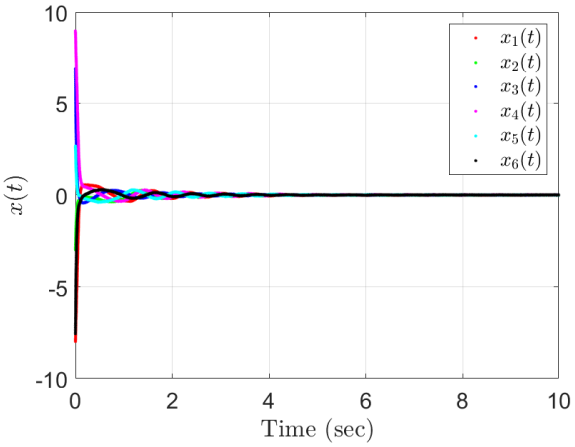

Figure 2. Concatenated state vector.

Table 2. RMS error values of state vector

State	Value
x_{1RMS}	0.434
x_{2RMS}	0.145
x_{3RMS}	0.233
x_{4RMS}	0.509
x_{5RMS}	0.134
x_{6RMS}	0.267

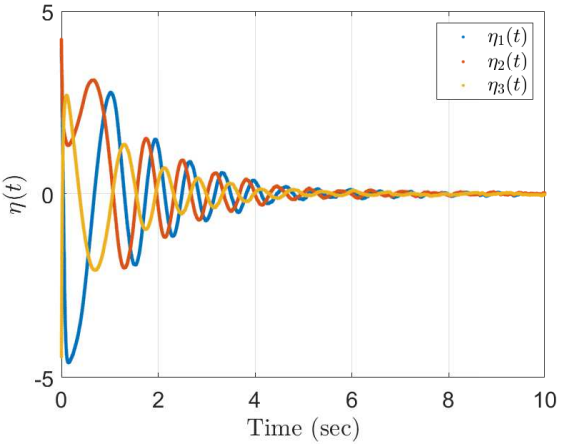

Figure 3. Position of the SS.

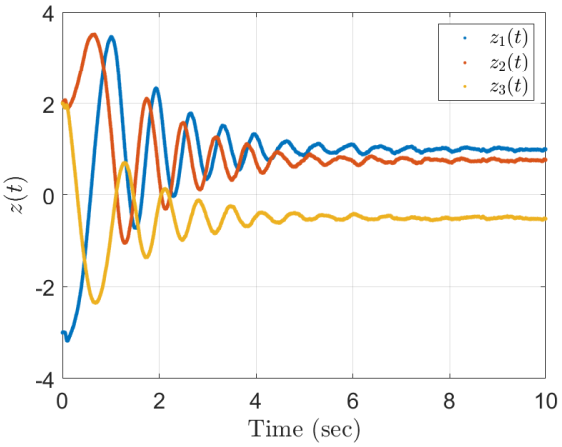

Figure 4. Position of the DS.

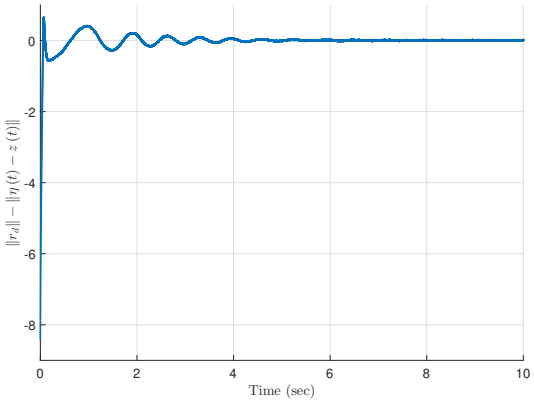

Figure 5. Error of the desired distance of the SS from the DS.

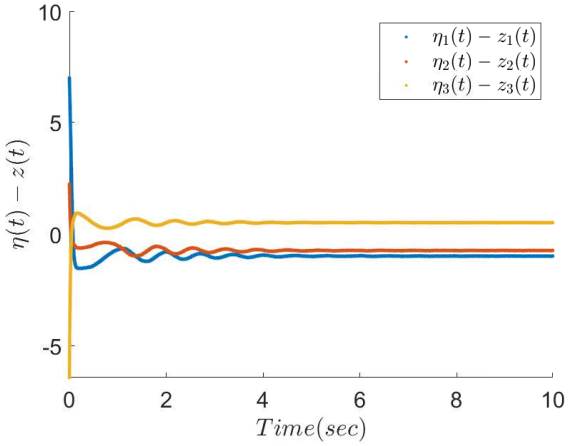

Figure 6. Difference between $\eta(t)$ and $z(t)$.

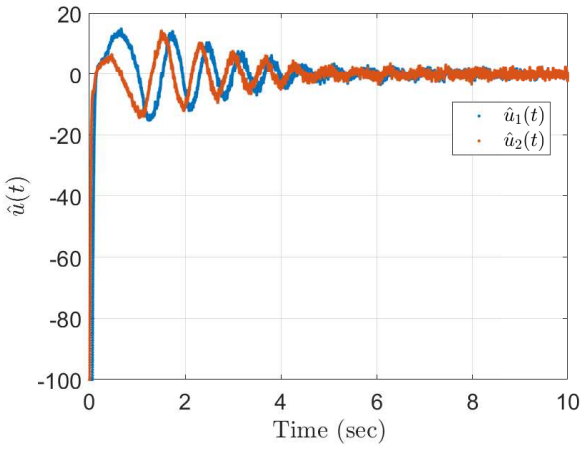

Figure 7. SS control policy.

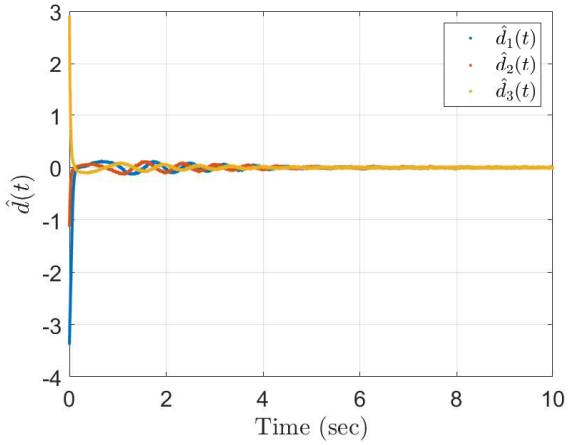

Figure 8. DS disturbing policy.

VI. Conclusion

A two-player non-cooperative zero-sum-game is formulated for a SS to intercept and regulate the orbital state of a DS. A StaF-based ADP method was used to approximate the unknown dynamics of the DS, including the uncertainty in the interaction forces with the SS, and the unknown value function. Reinforcement learning and a BE extrapolation method are used to identify the unknown dynamics provided a mild sufficient excitation condition is satisfied. A Lyapunov-based analysis is used to conclude uniformly ultimately bounded steady-state approximation and regulation errors. Initial simulations demonstrate the feasibility and performance of the developed controller under a set of simplifying assumptions. Additional efforts will target also including uncertainties in the SS dynamics, and including a more realistic Coulomb force model that may allow for the SS to exert pushing and pulling forces on the DS.

VII. Acknowledgments

This research is supported by AFOSR award number FA9550-19-1-0169, Office of Naval Research Grant N00014-13-1-0151, and NSF award number 1509516. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of sponsoring agencies.

References

[1]J.-C. Liou, D. Hall, P. Krisko, and J. Opiela, "LEGEND - a three-dimensional LEO-to-GEO debris evolutionary model," *Adv. Space Res.*, vol. 34, no. 5, pp. 981–986, 2004.

[2]Anon, "Space debris by the numbers," European Space Agency, Tech. Rep., 2017.

[3]——, "Report of the Scientific and Technical Subcommittee of the United Nations Committee on the Peaceful Uses of Outer Space on the 44th session," United Nations, Tech. Rep. A/AC.105/890, 2007.

[4]J. C. Liou and N. L. Johnson, "Risks in space from orbiting debris," *Science*, vol. 311, pp. 340–341, 2006.

[5]J.-C. Liou and N. Johnson, "Instability of the present LEO satellite populations," *Adv. Space Res.*, vol. 41, pp. 1046–1053, 2008.

[6]J. chyi Liou, "Active debris removal - a grand engineering challenge for the twenty-first century," 2010.

[7]Anon, "IADC space debris mitigation guidelines," Inter-Agency Space Debris Coordination Committee, Tech. Rep., 2007.

[8]C. P. Mark and S. Kamath, "Review of active space debris removal methods," *Space Policy*, vol. 47, pp. 194–206, 2019.

[9]M. Shan, J. Guo, and E. Gill, "Review and comparison of active space debris capturing and removal methods," *Prog. in Aerosp. Sci.*, vol. 80, pp. 18–32, 2016.

[10]J. Berryman and H. Schaub, "Analytical charge analysis for two-and three-craft coulomb formations," *AIAA J. Guid. Control Dyn.*, vol. 30, no. 6, pp. 1701–1710, 2007.

[11]H. Schaub and D. F. Moorer, "Geosynchronous large debris reorbiter: challenges and prospects," *The J. of the Astronaut. Sci.*, vol. 59, no. 1-2, pp. 161–176, 2012.

[12]K. Wormnes, R. Le Letty, L. Summerer, R. Schonenborg, O. Dubois-Matra, E. Luraschi, A. Cropp, H. Krag, and J. Delaval, "Esa technologies for space debris remediation," in *Eur. Conf. Sp. Debris*, vol. 1. ESA Communications ESTEC, Noordwijk, The Netherlands, 2013, pp. 1–8.

[13]L. B. King, G. G. Parker, S. Deshmukh, and J.-H. Chong, "Spacecraft formation-flying using inter-vehicle coulomb forces," *NIAC Ph. I Final Rep.*, 2002.

[14]H. Schaub, G. G. Parker, and L. B. King, "Challenges and prospects of coulomb spacecraft formation control," *J. Astronaut. Sci.*, vol. 52, no. 1, pp. 169–193, 2004.

[15]D. R. Jones and H. Schaub, "Periodic relative orbits of two spacecraft subject to differential gravity and electrostatic forcing," *Acta Astronaut.*, vol. 89, pp. 21 – 30, 2013.

[16]N. Zinner, A. Williamson, K. Brenner, J. Curran, A. Isaak, M. Knoch, A. Leppek, and J. Lestishen, "Junk hunter: Autonomous rendezvous, capture, and de-orbit of orbital debris," in *AIAA SPACE 2011 Conf. & Expos.*, 2011, p. 7292.

[17]C. Bombardelli and J. Pelaez, "Ion beam shepherd for contactless space debris removal," *AIAA J. Guid. Control Dyn.*, vol. 34, no. 3, pp. 916–920, 2011.

[18]P. Deptula, Z. I. Bell, F. Zegers, R. Licitra, and W. E. Dixon, "Single agent indirect herding via approximate dynamic programming," in *Proc. IEEE Conf. Decis. Control*, Dec. 2018, pp. 7136–7141.

[19]R. Isaacs, *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*, ser. Dover Books on Mathematics. Dover Publications, 1999.

[20]——, *Differential Games.* John Wiley, 1967.

[21]P. Deptula, J. Rosenfeld, R. Kamalapurkar, and W. E. Dixon, "Approximate dynamic programming: Combining regional and local state following approximations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2154–2166, June 2018.

[22]J. A. Rosenfeld, R. Kamalapurkar, and W. E. Dixon, "The state following (staf) approximation method," *IEEE Trans. on Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1716–1730, June 2019.

[23]R. Kamalapurkar, P. S. Walters, J. A. Rosenfeld, and W. E. Dixon, *Reinforcement learning for optimal feedback control: A Lyapunov-based approach.* Springer, 2018.

[24]E. Garcia, D. W. Casbeer, and M. Pachter, "Active target defence differential game: fast defender case," *IET Control Theory Appl.*, vol. 11, no. 17, pp. 2985–2993, 2017.

[25]——, "Design and analysis of state-feedback optimal strategies for the differential game of active defense," *IEEE Trans Autom. Contol*, 2018.

[26]M. V. Ramana and M. Kothari, "Pursuit-evasion games of high speed evader," *J. Intell. Rob. Syst.*, vol. 85, no. 2, pp. 293–306, 2017.

[27]M. Chen, Z. Zhou, and C. J. Tomlin, "Multiplayer reach-avoid games via pairwise outcomes," *IEEE Trans. Autom. Control*, vol. 62, no. 3, pp. 1451–1457, 2017.

[28]J. Chen, W. Zha, Z. Peng, and D. Gu, "Multi-player pursuit–evasion games with one superior evader," *Automatica*, vol. 71, pp. 24–32, 2016.

[29]K. G. Vamvoudakis and F. L. Lewis, "Multi-player non-zero-sum games: Online adaptive learning solution of coupled hamilton-jacobi equations," *Automatica*, vol. 47, pp. 1556–1569, 2011.

[30]R. Kamalapurkar, J. R. Klotz, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for differential graphical games," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 1, pp. 423–433, 2018.

[31]M. Johnson, R. Kamalapurkar, S. Bhasin, and W. E. Dixon, "Approximate n-player nonzero-sum game solution for an uncertain continuous nonlinear system," *IEEE Trans. on Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1645–1658, Aug. 2015.

[32]F. L. Lewis and D. Liu, *Reinforcement learning and approximate dynamic programming for feedback control.* John Wiley & Sons, 2013, vol. 17.

[33]D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming.* Athena Scientific, 1996.

[34]D. Bertsekas, "Approximate policy iteration: a survey and some new methods," *J. Control Theory Appl.*, vol. 9, pp. 310–335, 2011.

[35]S. S. Kumkov, S. Le Ménec, and V. S. Patsko, "Zero-sum pursuit-evasion differential games with many objects: survey of publications," *Dyn. Games Appl.*, vol. 7, no. 4, pp. 609–633, 2017.

[36]R. Vidal, O. Shakernia, H. Kim, D. Shim, and S. Sastry, "Probabilistic pursuit-evasion games: theory, implementation, and experimental evaluation," *IEEE Trans. Robot. and Autom.*, vol. 18, no. 5, pp. 662–669, Oct. 2002.

[37]S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin, "Hamilton-jacobi reachability: A brief overview and recent advances," in *Proc. IEEE Conf. Decis. Control.* IEEE, 2017, pp. 2242–2253.

[38]H. Zhu, Y.-Z. Luo, Z.-Y. Li, Z. Yang *et al.*, "Orbital pursuit-evasion games with incomplete information in the hill reference frame," in *Sens. and Syst. for Sp. Appl.* Engineers Australia, Royal Aeronautical Society., 2019, p. 1705.

[39]E. P. Blasch, K. Pham, and D. Shen, "Orbital satellite pursuit-evasion game-theoretical control," in *Int. Conf. on Inf. Sci., Signal Process. and their Appl.* IEEE, 2012, pp. 1007–1012.

[40]T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory: Second Edition*, ser. Classics in Applied Mathematics. SIAM, 1999.

[41]T. Basar and P. Bernhard, *Hinfinity- optimal control and related minimax design problems: A dynamic game approach.* Birkhäuser, 1995.

[42]Q. Jiao, H. Modares, S. Xu, F. L. Lewis, and K. G. Vamvoudakis, "Multi-agent zero-sum differential graphical games for disturbance rejection in distributed control," *Automatica*, vol. 69, pp. 24–34, 2016.

[43]F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*, 3rd ed. Hoboken, NJ: Wiley, 2012.

[44]S. Xue, B. Luo, and D. Liu, "Event-triggered adaptive dynamic programming for zero-sum game of partially unknown continuous-time nonlinear systems," *IEEE Trans. Syst. Man Cybern., Syst.*, no. 99, pp. 1–11, 2018.

[45]D. Wang, H. He, and D. Liu, "Improving the critic learning for event-based nonlinear H_∞ control design," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3417–3428, 2017.

[46]J. A. Farrell and M. M. Polycarpou, *Adaptive approximation based control: Unifying neural, fuzzy and traditional adaptive approximation approaches*, ser. Adaptive and Learning Systems for Signal Processing, Communications and Control Series. John Wiley & Sons, 2006, vol. 48.

[47]M. H. Stone, "The generalized weierstrass approximation theorem," *Math. Mag.*, vol. 21, no. 4, pp. 167–184, 1948.

[48]R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, Jan. 2015.

[49]R. Kamalapurkar, L. Andrews, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for infinite-horizon approximate optimal tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 753–758, 2017.

[50]R. Kamalapurkar, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for approximate optimal regulation," *Automatica*, vol. 64, pp. 94–104, 2016.

[51]R. Kamalapurkar, J. Rosenfeld, and W. E. Dixon, "Efficient model-based reinforcement learning for approximate online optimal control," *Automatica*, vol. 74, pp. 247–258, Dec. 2016.

[52]H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[53]J. A. Bittencourt, *Fundamentals of plasma physics.* Springer Sci. & Bus. Media, 2013.