Online Approximate Optimal Station Keeping of a Marine Craft in the Presence of an Irrotational Current

Patrick Walters, Rushikesh Kamalapurkar, *Senior Member, IEEE*, Forrest Voight, Eric M. Schwartz, *Senior Member, IEEE*, and Warren E. Dixon, *Fellow, IEEE*

***Abstract*—Online approximation of the optimal station-keeping strategy for a marine craft subject to an irrotational current is considered. An approximate policy that minimizes a user-defined cost function over an infinite time horizon is obtained using an actor-critic-identifier-based adaptive dynamic programming technique. The hydrodynamic drift dynamics are assumed to be unknown; therefore, a concurrent learning-based system identifier is developed to identify the unknown model parameters. The identified model is used to implement an adaptive model-based reinforcement learning technique to estimate the unknown value function. The developed policy guarantees uniformly ultimately bounded convergence of the vehicle to the desired station and uniformly ultimately bounded convergence of the approximated policies to the optimal polices without the requirement of persistence of excitation. The developed strategy is validated using an autonomous underwater vehicle, where the three degrees-of-freedom in the horizontal plane are regulated. The experiments are conducted in a second-magnitude spring located in central Florida.**

***Index Terms*—Adaptive dynamic programming (ADP), marine craft, nonlinear control, station keeping.**

I. Introduction

MARINE craft, which include ships, floating platforms, autonomous underwater vehicles (AUVs), etc., play a vital role in commercial, military, and recreational objectives. Marine craft are often required to remain on a station for an extended period of time, e.g., floating oil platforms, support vessels, and AUVs acting as a communication link for multiple vehicles or persistent environmental monitors. The success of the vehicle often relies on the vehicle's ability to hold a precise station (e.g., station keeping near structures or underwater features). The cost of holding that station is correlated with the energy expended for propulsion through consumption of fuel and wear on mechanical systems, especially when station keeping in environments with a persistent current. Therefore, by reducing the energy expended for station-keeping objectives, the cost of holding a station can be reduced.

Precise station keeping of a marine craft is challenging because of nonlinearities in the dynamics of the vehicle. A survey of station keeping for surface vessels can be found in [1]. Common approaches employed to control a marine craft include robust and adaptive control methods [2]–[5]. These methods provide robustness to disturbances and/or model uncertainty; however, they do not explicitly account for the cost of the control effort. Motivated by the desire to balance energy expenditure and the accuracy of the vehicle' s station, approximate optimal control methods are examined in this paper to minimize a user-defined cost function of the total control effort (energy expended) and state error (station accuracy). Because of the difficulties associated with finding closed form analytical solutions to optimal control problems for marine craft, efforts such as [6] numerically approximate the solution to the Hamilton–Jacobi–Bellman (HJB) equation using an iterative application of Galerkin's method.

Various methods have been proposed to find an approximate solution to the HJB equation. Adaptive dynamic programming (ADP) is one such method in which a solution to the HJB equation is approximated using parametric function approximation techniques. ADP-based techniques have been used to approximate optimal control policies for regulation (e.g., [7]–[12]) of general nonlinear systems. Efforts in [13] and [14] present ADP-based solutions to the Hamilton–Jacobi–Isaacs equation that yield an approximate optimal policy accounting for state-dependent disturbances. However, these methods do not consider explicit time-varying disturbances such as the dynamics that are introduced due to the presence of current.

In this result, an optimal station-keeping policy that captures the desire to balance the need to accurately hold a station and

Manuscript received August 18, 2017; revised November 16, 2017; accepted November 29, 2017. Date of publication March 29, 2018; date of current version April 12, 2018. This paper was recommended for publication by Associate Editor F. Boyer and Editor A. Billard upon evaluation of the reviewers' comments. This work was supported in part by National Science Foundation under Grant 0901491, Grant 1161260, and Grant 1217908, in part by Office of Naval Research under Grant N00014-13-1-0151, and in part by a contract with the Air Force Research Laboratory Mathematical Modeling and Optimization Institute. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agency. *(Corresponding author: Patrick Walters.)*

P. Walters and W. E. Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: walters8@ufl.edu; wdixon@ufl.edu).

R. Kamalapurkar is with the Department of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK 74074 USA (e-mail: rushikesh.kamalapurkar@okstate.edu).

F. Voight and E. M. Schwartz are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: forrestv@ufl.edu; ems@ufl.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2018.2791600

the cost of holding that station through a user-defined quadratic performance criterion is generated for a fully actuated marine craft. The developed controller differs from results such as [8] and [9] in that it tackles the challenges associated with the introduction of a time-varying irrotational current. A complicating factor in the development is that the presence of an irrotational time-varying current makes the system nonautonomous. To account for this aspect, an approximate optimal station-keeping controller is developed for a residual (disturbance free) model, where the effects of the irrotational current and other differences between the residual model and the actual dynamics are included in the stability analysis. An extension to include the current if it is constant is included in Appendix A. Since the hydrodynamic parameters of a marine craft are often difficult to determine, a concurrent learning system identifier is developed. As outlined in [15], concurrent learning uses additional information from recorded data to remove the persistence of excitation requirement associated with traditional system identifiers. The proposed model-based ADP method generates the optimal station-keeping policy using a combination of on-policy and off-policy data, eliminating the need for physical exploration of the state space. A Lyapunov-based stability analysis is presented which guarantees uniformly ultimately bounded (UUB) convergence of the marine craft to its station and UUB convergence of the approximated policy to the optimal policy.

To illustrate the performance of the developed controller, an AUV is used to collect experimental data. Specifically, the developed strategy is implemented for planar regulation of an AUV near the vent of a second-magnitude spring located in central Florida. The experimental results demonstrate the developed method' s ability to simultaneously identify the unknown hydrodynamic parameters and generate an approximate optimal policy using the identified model in the presence of a current.

II. Vehicle Model

Consider the nonlinear equations of motion for a marine craft including the effects of irrotational current given in [16, Sec. 7.5] as

$$\dot{\eta} = J_E(\eta)\,\nu \tag{1}$$

$$M_{RB}\dot{\nu} + C_{RB}(\nu)\,\nu + M_A\dot{\nu}_r + C_A(\nu_r)\,\nu_r + D_A(\nu_r)\,\nu_r + G(\eta) = \tau_b \tag{2}$$

where $\nu \in \mathbb{R}^n$ is the body-fixed translational and angular velocity vector, $\nu_c \in \mathbb{R}^n$ is the body-fixed irrotational current velocity vector, $\nu_r = \nu - \nu_c$ is the relative body-fixed translational and angular fluid velocity vector, $\eta \in \mathbb{R}^n$ is the earth-fixed position and orientation vector, $J_E : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is the coordinate transformation between the body-fixed and earth-fixed coordinates,[1] $M_{RB} \in \mathbb{R}^{n\times n}$ is the constant rigid body inertia matrix, $C_{RB} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is the rigid body centripetal and Coriolis matrix, $M_A \in \mathbb{R}^{n\times n}$ is the constant hydrodynamic added mass matrix, $C_A : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is the unknown hydrodynamic centripetal and Coriolis matrix, $D_A : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is the unknown hydrodynamic damping and friction matrix, $G : \mathbb{R}^n \to \mathbb{R}^n$ is the gravitational and buoyancy force and moment vector, and $\tau_b \in \mathbb{R}^n$ is the body-fixed force and moment control input.

In the case of a three degree-of-freedom (DOF) planar model with orientation represented as Euler angles, the state vectors in (1) and (2) are further defined as

$$\eta \triangleq \begin{bmatrix} x & y & \psi \end{bmatrix}^T$$

$$\nu \triangleq \begin{bmatrix} u_b & v_b & r_b \end{bmatrix}^T$$

where $x, y \in \mathbb{R}$, are the earth-fixed position vector components of the center of mass, $\psi \in [0, 2\pi]$ represents the yaw angle, u_b, $v_b \in \mathbb{R}$ are the body-fixed translational velocities, and $r_b \in \mathbb{R}$ is the body-fixed angular velocity. The irrotational current vector is defined as

$$\nu_c \triangleq \begin{bmatrix} u_c & v_c & 0 \end{bmatrix}^T$$

where $u_c, v_c \in \mathbb{R}$ are the body-fixed current translational velocities. The coordinate transformation $J_E(\eta)$ is given as

$$J_E(\eta) = \begin{bmatrix} \cos(\psi) & -\sin(\psi) & 0 \\ \sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Assumption 1: The marine craft is neutrally buoyant if submerged and the center of gravity is located vertically below the center of buoyancy on the z-axis if the vehicle model includes roll and pitch.[2]

III. System Identifier

Since the hydrodynamic effects pertaining to a specific marine craft may be unknown, an online system identifier is developed for the vehicle drift dynamics. Consider the control affine form of the vehicle model

$$\dot{\zeta} = Y(\zeta, \nu_c)\,\theta + f_0(\zeta, \dot{\nu}_c) + g\tau_b \tag{3}$$

where $\zeta \triangleq \begin{bmatrix} \eta & \nu \end{bmatrix}^T \in \mathbb{R}^{2n}$ is the state vector. The unknown hydrodynamics are linear-in-the-parameters with p unknown parameters where $Y : \mathbb{R}^{2n} \times \mathbb{R}^n \to \mathbb{R}^{2n\times p}$ is the regression matrix and $\theta \in \mathbb{R}^p$ is the vector of unknown parameters. The unknown hydrodynamic effects are modeled as[3]

$$Y(\zeta, \nu_c)\,\theta = \begin{bmatrix} 0 \\ -M^{-1}C_A(\nu_r)\,\nu_r - M^{-1}D_A(\nu_r)\,\nu_r \end{bmatrix}$$

and known rigid body drift dynamics $f_0 : \mathbb{R}^{2n} \times \mathbb{R}^n \to \mathbb{R}^{2n}$ are modeled as

$$f_0(\zeta, \dot{\nu}_c) = \begin{bmatrix} J_E(\eta)\,\nu \\ M^{-1}M_A\dot{\nu}_c - M^{-1}C_{RB}(\nu)\,\nu - M^{-1}G(\eta) \end{bmatrix}$$

[1]The orientation of the vehicle may be represented as Euler angles, quaternions, or angular rates. In this development, the use of Euler angles is assumed, see [16, Sec. 7.5] for details regarding other representations.

[2]This assumption simplifies the subsequent analysis and can often be met by trimming the vehicle. For marine craft where this assumption cannot be met, an additional term may be added to the controller, similar to how terms dependent on the irrotational current are handled.

[3]Instead of the assuming linear in the parameters assumption, the dynamics could also be approximated using a neural network (NN), for example, where the ideal weights could have been approximated by a concurrent learning-based update law.

where $M \triangleq M_{RB} + M_A$, and the body-fixed current velocity ν_c, and acceleration $\dot{\nu}_c$ are assumed to be measurable.[4] The known constant control effectiveness matrix $g \in \mathbb{R}^{2n\times n}$ is defined as

$$g \triangleq \begin{bmatrix} 0 \\ M^{-1} \end{bmatrix}.$$

An identifier is designed as

$$\dot{\hat{\zeta}} = Y(\zeta, \nu_c)\hat{\theta} + f_0(\zeta, \dot{\nu}_c) + g\tau_b + k_\zeta \tilde{\zeta} \tag{4}$$

where $\tilde{\zeta} \triangleq \zeta - \hat{\zeta}$ is the measurable state estimation error, and $k_\zeta \in \mathbb{R}^{2n\times 2n}$ is a constant positive definite diagonal gain matrix. Subtracting (4) from (3) yields

$$\dot{\tilde{\zeta}} = Y(\zeta, \nu_c)\tilde{\theta} - k_\zeta \tilde{\zeta}$$

where $\tilde{\theta} \triangleq \theta - \hat{\theta}$ is the parameter identification error.

A. Parameter Update

Traditional adaptive control techniques require persistence of excitation to ensure the parameter estimates $\hat{\theta}$ converge to their true values θ (cf., [17] and [18]). Persistence of excitation often requires an excitation signal to be applied to the vehicle's input resulting in unwanted deviations in the vehicle state. These deviations are often in opposition to the vehicle's control objectives. Alternatively, a concurrent learning-based system identifier can be developed (cf., [15] and [19]). The concurrent learning-based system identifier relaxes the persistence of excitation requirement through the use of a prerecorded history stack of state-action pairs.[5]

Assumption 2: There exists a prerecorded dataset of sampled data points $\{\zeta_j, \nu_{cj}, \dot{\nu}_{cj}, \tau_{bj} \in \chi | j = 1, 2, \ldots, M\}$ with a numerically calculated state derivatives $\dot{\bar{\zeta}}_j$ at each recorded state-action pair such that $\forall t \in [0, \infty)$

$$\operatorname{rank}\left(\sum_{j=1}^{M} Y_j^T Y_j\right) = p \tag{5}$$

$$\left\|\dot{\bar{\zeta}}_j - \dot{\zeta}_j\right\| < \bar{d}, \forall j$$

where $Y_j \triangleq Y(\zeta_j, \nu_{cj})$, $f_{0j} \triangleq f_0(\zeta_j)$, $\dot{\zeta}_j = Y_j\theta + f_{0j} + g\tau_{bj}$, and $\bar{d} \in [0, \infty)$ is a constant.

[4]The body-fixed current velocity ν_c may be trivially measured using sensors commonly found on marine craft, such as a Doppler velocity log (DVL), while the current acceleration $\dot{\nu}_c$ may be determined using numerical differentiation and smoothing.

[5]In this development, it is assumed that a dataset of state-action pairs is available *a priori*. Experiments can be done to collect the state-action pair data; however, they do not necessarily need to be conducted in the presence of a current (e.g., the data may be collected in a pool). Since the current effects the dynamics only through the ν_r terms, data that are sufficiently rich and satisfies Assumption 2 may be collected by merely exploring the ζ state space. Note, this is the reason the body-fixed current ν_c and acceleration $\dot{\nu}_c$ are not considered a part of the state. If state-action data are not available for the given system then it is possible to build the history stack in real time and the details of that development can be found in [20, Appendix A].

The parameter estimate update law is given as

$$\dot{\hat{\theta}} = \Gamma_\theta Y(\zeta, \nu_c)^T \tilde{\zeta} + \Gamma_\theta k_\theta \sum_{j=1}^{M} Y_j^T \left(\dot{\bar{\zeta}}_j - f_{0j} - g\tau_{bj} - Y_j\hat{\theta}\right) \tag{6}$$

where Γ_θ is a positive definite, diagonal gain matrix, and k_θ is a positive, scalar gain matrix. To facilitate the stability analysis, the parameter estimate update law is expressed in the advantageous form

$$\dot{\hat{\theta}} = \Gamma_\theta Y(\zeta, \nu_c)^T \tilde{\zeta} + \Gamma_\theta k_\theta \sum_{j=1}^{M} Y_j^T \left(Y_j\tilde{\theta} + d_j\right)$$

where $d_j = \dot{\bar{\zeta}}_j - \dot{\zeta}_j$.

Remark 1: The update law in (6) does not require instantaneous measurement of acceleration. Acceleration only needs to be computed at the past time instances when the data points $(\zeta_j, \nu_{cj}, \dot{\nu}_{cj}, \tau_{bj})$ were recorded. Acceleration at a past time instance t^* can be accurately computed by recording position and velocity signals over a time interval that contains t^* in its interior and using noncausal estimation methods such as optimal fixed-point smoothing [21, p. 170].

B. Convergence Analysis

Consider the candidate Lyapunov function $V_P : \mathbb{R}^{2n+p} \times [0, \infty)$ given as

$$V_P(Z_P) = \frac{1}{2}\tilde{\zeta}^T\tilde{\zeta} + \frac{1}{2}\tilde{\theta}^T\Gamma_\theta^{-1}\tilde{\theta} \tag{7}$$

where $Z_P \triangleq \begin{bmatrix} \tilde{\zeta}^T & \tilde{\theta}^T \end{bmatrix}$. The candidate Lyapunov function can be bounded as

$$\frac{1}{2}\min\left\{1, \underline{\gamma_\theta}\right\}\|Z_P\|^2 \leq V_P(Z_P) \leq \frac{1}{2}\max\left\{1, \overline{\gamma_\theta}\right\}\|Z_P\|^2 \tag{8}$$

where $\underline{\gamma_\theta}, \overline{\gamma_\theta}$ are the minimum and maximum eigenvalues of Γ_θ, respectively.

The time derivative of the candidate Lyapunov function in (7) is

$$\dot{V}_P = -\tilde{\zeta}^T k_\zeta \tilde{\zeta} - k_\theta \tilde{\theta}^T \sum_{j=1}^{M} Y_j^T Y_j \tilde{\theta} - k_\theta \tilde{\theta}^T \sum_{j=1}^{M} Y_j^T d_j.$$

The time derivative may be upper bounded by

$$\dot{V}_P \leq -\underline{k_\zeta}\left\|\tilde{\zeta}\right\|^2 - k_\theta \underline{y}\left\|\tilde{\theta}\right\|^2 + k_\theta d_\theta \left\|\tilde{\theta}\right\| \tag{9}$$

where $\underline{k_\zeta}, \underline{y}$ are the minimum eigenvalues of k_ζ and $\sum_{j=1}^{M} Y_j^T Y_j$, respectively, and $d_\theta = \bar{d}\sum_{j=1}^{M}\|Y_j\|$. Completing the squares, (9) may be upper bounded by

$$\dot{V}_P \leq -\underline{k_\zeta}\left\|\tilde{\zeta}\right\|^2 - \frac{k_\theta \underline{y}}{2}\left\|\tilde{\theta}\right\|^2 + \frac{k_\theta d_\theta^2}{2\underline{y}}$$

which may be further upper bounded by

$$\dot{V}_P \leq -\alpha_P\|Z_P\|^2, \forall \|Z_P\| \geq K_P > 0 \tag{10}$$

where $\alpha_P \triangleq \frac{1}{2}\min\left\{2\underline{k_\zeta}, k_\theta \underline{y}\right\}$ and $K_P \triangleq \sqrt{\frac{k_\theta d_\theta^2}{2\alpha_P \underline{y}}}$. Using (8) and (10), $\tilde{\zeta}$ and $\tilde{\theta}$ can be shown to exponentially decay to a ultimate bound as $t \to \infty$. The ultimate bound may be made arbitrarily small depending on the selection of the gains k_ζ and k_θ.

IV. Problem Formulation

A. Residual Model

The presence of a time-varying irrotational current yields unique challenges in the formulation of the optimal regulation problem. Since the current renders the system nonautonomous, a residual model that does not include the effects of the irrotational current is introduced. The residual model is used in the development of the optimal control problem in place of the original model. A disadvantage of this approach is that the optimal policy is developed for the current-free model.[6] In the case where the earth-fixed current is constant, the effects of the current may be included in the development of the optimal control problem as detailed in Appendix A.

The residual model can be written in a control affine form as

$$\dot{\zeta} = Y_{\text{res}}(\zeta)\,\theta + f_{0_{\text{res}}}(\zeta) + gu \tag{11}$$

where the unknown hydrodynamics are linear-in-the-parameters with p unknown parameters where $Y_{\text{res}} : \mathbb{R}^{2n} \to \mathbb{R}^{2n\times p}$ is a regression matrix, the function $f_{0_{\text{res}}} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is the known portion of the dynamics, and $u \in \mathbb{R}^n$ is the control vector. The drift dynamics, defined as $f_{\text{res}}(\zeta) = Y_{\text{res}}(\zeta)\,\theta + f_{0_{\text{res}}}(\zeta)$, can be shown to satisfy $f_{\text{res}}(0) = 0$ when Assumption 1 is satisfied.

The drift dynamics in (11) are modeled as

$$Y_{\text{res}}(\zeta)\,\theta = \begin{bmatrix} 0 \\ -M^{-1}C_A(\nu)\,\nu - M^{-1}D_A(\nu)\,\nu \end{bmatrix}$$

$$f_{0_{\text{res}}}(\zeta) = \begin{bmatrix} J_E\nu \\ -M^{-1}C_{RB}(\nu)\,\nu - M^{-1}G(\eta) \end{bmatrix} \tag{12}$$

and the virtual control vector u is defined as

$$u = \tau_b - \tau_c(\zeta, \nu_c, \dot{\nu}_c) \tag{13}$$

where $\tau_c : \mathbb{R}^{2n} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a feedforward term to compensate the effect of the variable current, which includes cross-terms generated by the introduction of the residual dynamics and is given as

$$\tau_c(\zeta, \nu_c, \dot{\nu}_c) = C_A(\nu_r)\,\nu_r + D_A(\nu_r)\,\nu_r - M_A\dot{\nu}_c - C_A(\nu)\,\nu - D_A(\nu)\,\nu.$$

The current feedforward term is represented in the advantageous form

$$\tau_c(\zeta, \nu_c, \dot{\nu}_c) = -M_A\dot{\nu}_c + Y_c(\zeta, \nu_c)\,\theta$$

where $Y_c : \mathbb{R}^{2n} \times \mathbb{R}^n \to \mathbb{R}^{2n\times p}$ is the regression matrix and

$$Y_c\theta(\zeta, \nu_c) = C_A(\nu_r)\,\nu_r + D_A(\nu_r)\,\nu_r - C_A(\nu)\,\nu - D_A(\nu)\,\nu.$$

Since the parameters are unknown, an approximation of the compensation term τ_c given by

$$\hat{\tau}_c\left(\zeta, \nu_c, \dot{\nu}_c, \hat{\theta}\right) = -M_A\dot{\nu}_c + Y_c\hat{\theta} \tag{14}$$

is implemented, and the approximation error is defined by

$$\tilde{\tau}_c \triangleq \tau_c - \hat{\tau}_c.$$

B. Nonlinear Optimal Regulation Problem

This section provides a review of the traditional infinite horizon optimal regulation problem to facilitate the subsequent development. The performance index for the optimal regulation problem is selected as

$$J(\zeta, u) = \int_0^\infty r(\zeta(\tau), u(\tau))\,d\tau \tag{15}$$

where $r : \mathbb{R}^{2n} \to [0, \infty)$ is the local cost defined as

$$r(\zeta, u) \triangleq \zeta^T Q\zeta + u^T Ru. \tag{16}$$

In (16), $Q \in \mathbb{R}^{2n\times 2n}$, $R \in \mathbb{R}^{n\times n}$ are symmetric positive definite weighting matrices, and u is the virtual control vector. The matrix Q has the property $\underline{q}\|\xi_q\|^2 \le \xi_q^T Q\xi_q \le \overline{q}\|\xi_q\|^2\ \forall \xi_q \in \mathbb{R}^{2n}$ where $\underline{q}$ and $\overline{q}$ are positive constants. The infinite-time scalar value function $V : \mathbb{R}^{2n} \to [0, \infty)$ for the optimal solution is written as

$$V(\zeta) = \min_u \int_0^\infty r(\zeta(\tau), u(\tau))\,d\tau. \tag{17}$$

The objective of the optimal control problem is to find the optimal policy $u^* : \mathbb{R}^{2n} \to \mathbb{R}^n$ that minimizes the performance index (15) subject to the dynamic constraints in (11). Assuming that a minimizing policy exists and the value function is continuously differentiable, the Hamiltonian $H : \mathbb{R}^{2n} \to \mathbb{R}$ is defined as

$$H(\zeta) \triangleq r(\zeta, u^*(\zeta)) + \frac{\partial V(\zeta)}{\partial \zeta}\left(Y_{\text{res}}(\zeta)\,\theta + f_{0_{\text{res}}}(\zeta) + gu^*(\zeta)\right). \tag{18}$$

The HJB equation is given as [22]

$$0 = \frac{\partial V(\zeta)}{\partial t} + H(\zeta) \tag{19}$$

where $\frac{\partial V(\zeta)}{\partial t} = 0$ since the value function is not an explicit function of time. After substituting (16) into (19), the optimal policy is given by [22]

$$u^*(\zeta) = -\frac{1}{2}R^{-1}g^T\left(\frac{\partial V(\zeta)}{\partial \zeta}\right)^T. \tag{20}$$

The analytical expression for the optimal controller in (20) requires knowledge of the value function which is the solution to the HJB equation in (19). The HJB equation is a partial differential equation which is generally infeasible to solve; hence, an approximate solution is sought.

[6] To the author's knowledge, there is no method to generate a policy with time-varying inputs (e.g., time-varying irrotational current) that guarantees optimally and stability.

V. Approximate Policy

The subsequent development is based on a NN approximation of the value function and optimal policy. Differing from previous ADP literature with model uncertainty (e.g., [9], [11], [12]) that seeks a NN approximation using the integral form of the HJB, the following development seeks a NN approximation using the differential form of the HJB, similar to [8]. In contrast to [8] where model identification was achieved using a dynamic NN using a robust integral of the sign of the error (RISE) feedback term [23], the differential form of the HJB coupled with the identified model via concurrent learning allows for off-policy learning, through Bellman error extrapolation.

Over any compact domain $\chi \subset \mathbb{R}^{2n}$, the value function $V : \mathbb{R}^{2n} \to [0, \infty)$ can be represented by a single-layer NN with l neurons as

$$V(\zeta) = W^T \sigma(\zeta) + \epsilon(\zeta) \tag{21}$$

where $W \in \mathbb{R}^l$ is the ideal weight vector bounded above by a known positive constant, $\sigma : \mathbb{R}^{2n} \to \mathbb{R}^l$ is a bounded, continuously differentiable activation function, and $\epsilon : \mathbb{R}^{2n} \to \mathbb{R}$ is the bounded, continuously differential function reconstruction error. Using (20) and (21), the optimal policy can be represented by

$$u^*(\zeta) = -\frac{1}{2} R^{-1} g^T \left(\sigma'(\zeta)^T W + \epsilon'(\zeta)^T \right) \tag{22}$$

where $\sigma' : \mathbb{R}^{2n} \to \mathbb{R}^{l \times 2n}$ and $\epsilon' : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ are derivatives with respect to the state. Based on (21) and (22), NN approximations of the value function and the optimal policy are defined as

$$\hat{V}\left(\zeta, \hat{W}_c\right) = \hat{W}_c^T \sigma(\zeta) \tag{23}$$

$$\hat{u}\left(\zeta, \hat{W}_a\right) = -\frac{1}{2} R^{-1} g^T \sigma'(\zeta)^T \hat{W}_a \tag{24}$$

where $\hat{W}_c$, $\hat{W}_a \in \mathbb{R}^l$ are estimates of the constant ideal weight vector W. The weight estimation errors are defined as $\tilde{W}_c \triangleq W - \hat{W}_c$ and $\tilde{W}_a \triangleq W - \hat{W}_a$.

Substituting (11), (23), and (24) into (18), the approximate Hamiltonian $\hat{H} : \mathbb{R}^{2n} \times \mathbb{R}^p \times \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}$ is given as

$$\hat{H}\left(\zeta, \hat{\theta}, \hat{W}_c, \hat{W}_a\right) = r\left(\zeta, \hat{u}\left(\zeta, \hat{W}_a\right)\right) + \frac{\partial \hat{V}\left(\zeta, \hat{W}_c\right)}{\partial \zeta}\left(Y_{\text{res}}(\zeta)\hat{\theta} + f_{0_{\text{res}}}(\zeta) + g\hat{u}\left(\zeta, \hat{W}_a\right)\right). \tag{25}$$

The error between the optimal and approximate Hamiltonian is called the Bellman error $\delta : \mathbb{R}^{2n} \times \mathbb{R}^p \times \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}$, given as

$$\delta\left(\zeta, \hat{\theta}, \hat{W}_c, \hat{W}_a\right) = \hat{H}\left(\zeta, \hat{\theta}, \hat{W}_c, \hat{W}_a\right) - H(\zeta) \tag{26}$$

where $H(\zeta) = 0 \, \forall \zeta \in \mathbb{R}^{2n}$. Therefore, the Bellman error can be written in a measurable form as

$$\delta\left(\zeta, \hat{\theta}, \hat{W}_c, \hat{W}_a\right) = r\left(\zeta, \hat{u}\left(\zeta, \hat{W}_a\right)\right) + \hat{W}_c^T \omega\left(\zeta, \hat{\theta}, \hat{W}_a\right)$$

where $\omega : \mathbb{R}^{2n} \to \mathbb{R}^l$ is given by

$$\omega\left(\zeta, \hat{\theta}, \hat{W}_a\right) = \sigma'\left(Y_{\text{res}}(\zeta)\hat{\theta} + f_{0_{\text{res}}}(\zeta) + g\hat{u}\left(\zeta, \hat{W}_a\right)\right).$$

The Bellman error may be extrapolated to unexplored regions of the state space since it depends solely on the approximated system model and current NN weight estimates. In Section VI, Bellman error extrapolation is employed to establish UUB convergence of the approximate policy to the optimal policy without requiring persistence of excitation provided the following assumption is satisfied.[7]

Assumption 3: [24] There exists a positive constant $\underline{c}$ and set of states $\{\zeta_k \in \chi | k = 1, 2, \ldots, N\}$ such that

$$\inf_{t \in [0, \infty)} \left[\lambda_{\min} \left(\sum_{k=1}^{N} \frac{\omega_k \omega_k^T}{\rho_k} \right) \right] = \underline{c} \tag{27}$$

where $\omega_k \triangleq \omega\left(\zeta_k, \hat{\theta}, \hat{W}_a\right)$ and $\rho_k \triangleq 1 + k_\rho \omega_k{}^T \Gamma \omega_k$.

PE is generally required for online system identification and for the identification of the value function. In online system identification, data can be recorded while the system is excited and then used to drive the adaptation when the system is not excited. Hence, storage and reuse of data soften the PE requirement in online system identification to a *finite excitation* requirement. The estimated system model is used to soften the PE requirement in value function identification via Bellman error extrapolation. Unlike PE, which requires the regressor ω evaluated along the state trajectory to be linearly independent, on average, for all t, the condition in Assumption 3 is a spatial condition that requires the values of the regressor evaluated at several arbitrarily selected points in the state-space to be linearly independent, for all t. That is, simulated off-trajectory experience can be simultaneously utilized for learning. Hence, unlike PE, Assumption 3 is independent of the state trajectory of the real system. Furthermore, the PE requirement, by definition, is in a direct conflict with the station-keeping objective. That is, when the marine craft is stationary PE does not hold, whereas Assumption 3 can still be satisfied. More importantly, unlike PE, Assumption 3 can be met without adding a potentially destabilizing ad hoc probing signal to the controller.

The value function least squares update law based on minimization of the Bellman error is given by

$$\dot{\hat{W}}_c = -\Gamma \left(k_{c1} \frac{\omega\left(\zeta, \hat{\theta}, \hat{W}_a\right)}{\rho} \delta\left(\zeta, \hat{\theta}, \hat{W}_c, \hat{W}_a\right) + \frac{k_{c2}}{N} \sum_{k=1}^{N} \frac{\omega_k}{\rho_k} \delta_k \right) \tag{28}$$

$$\dot{\Gamma} = \begin{cases} \beta\Gamma - k_{c1}\Gamma \dfrac{\omega\left(\zeta, \hat{\theta}, \hat{W}_a\right)\omega\left(\zeta, \hat{\theta}, \hat{W}_a\right)^T}{\rho}\Gamma, & \|\Gamma\| \leq \overline{\Gamma} \\ 0, & \text{otherwise} \end{cases} \tag{29}$$

where k_{c1}, $k_{c2} \in \mathbb{R}$ are a positive adaptation gains, $\delta_k \triangleq \delta\left(\zeta_k, \hat{\theta}, \hat{W}_c, \hat{W}_a\right)$ is the extrapolated Bellman error, $\|\Gamma(t_0)\| =$

[7] Assumption 3 is used in the subsequent stability analysis to conclude the uncertain parameters are identified, in a similar, but less restrictive, manner as the traditional persistence of excitation condition.

$\|\Gamma_0\| \leq \bar{\Gamma}$ is the initial adaptation gain, $\bar{\Gamma} \in \mathbb{R}$ is a positive saturation gain, $\beta \in \mathbb{R}$ is a positive forgetting factor, and

$$\rho \triangleq 1 + k_\rho \omega \left(\zeta, \hat{\theta}, \hat{W}_a\right)^T \Gamma \omega \left(\zeta, \hat{\theta}, \hat{W}_a\right)$$

is a normalization constant, where $k_\rho \in \mathbb{R}$ is a positive gain. The update law in (28) and (29) ensures that

$$\underline{\Gamma} \leq \|\Gamma\| \leq \overline{\Gamma} \quad \forall t \in [0, \infty).$$

The actor NN update law is given by

$$\dot{\hat{W}}_a = \text{proj}\left\{-k_a \left(\hat{W}_a - \hat{W}_c\right)\right\} \tag{30}$$

where $k_a \in \mathbb{R}$ is an positive gain, and proj $\{\cdot\}$ is a smooth projection operator[8] used to bound the weight estimates. Using properties of the projection operator, the actor NN weight estimation error can be bounded above by positive constant.

Using the definition in (13), the force and moment applied to the vehicle, described in (3), is given in terms of the approximated optimal virtual control (24) and the compensation term approximation in (14) as

$$\hat{\tau}_b = \hat{u}\left(\zeta, \hat{W}_a\right) + \hat{\tau}_c\left(\zeta, \hat{\theta}, \nu_c, \dot{\nu}_c\right). \tag{31}$$

VI. Stability Analysis

For notational brevity, all function dependencies from previous sections will be henceforth suppressed. An unmeasurable form of the Bellman error can be written using (18), (25), and (26) as

$$\begin{aligned}\delta = & -\tilde{W}_c^T \omega - W^T \sigma' Y_{\text{res}} \tilde{\theta} - \epsilon'\left(Y_{\text{res}}\theta + f_{0_{\text{res}}}\right) \\ & + \frac{1}{4}\tilde{W}_a^T G_\sigma \tilde{W}_a + \frac{1}{2}\epsilon' G \sigma'^T W + \frac{1}{4}\epsilon' G \epsilon'^T\end{aligned} \tag{32}$$

where $G \triangleq g R^{-1} g^T \in \mathbb{R}^{2n \times 2n}$ and $G_\sigma \triangleq \sigma' G \sigma'^T \in \mathbb{R}^{l \times l}$ are symmetric, positive semidefinite matrices. Similarly, the Bellman error at the sampled data points can be written as

$$\delta_k = -\tilde{W}_c^T \omega_k - W^T \sigma_k' \left(Y_{\text{res}_k} \tilde{\theta}\right) + \frac{1}{4}\tilde{W}_a^T G_{\sigma k} \tilde{W}_a + E_k \tag{33}$$

where

$$E_k \triangleq \frac{1}{2}\epsilon_k' G \sigma_k'^T W + \frac{1}{4}\epsilon_k' G \epsilon_k'^T - \epsilon_k'\left(Y_{\text{res}_k}\theta + f_{0_{\text{res}_k}}\right) \in \mathbb{R}$$

is a constant at each data point, and the notation F_k denotes the function $F(\zeta, \cdot)$ evaluated at the sampled state, i.e., $F_k(\cdot) = F(\zeta_k, \cdot)$. The functions Y_{res} and $f_{0_{\text{res}}}$ on the compact set χ are Lipschitz continuous and can be bounded by

$$\begin{aligned}\|Y_{\text{res}}\| &\leq L_{Y_{\text{res}}} \|\zeta\| \quad \forall \zeta \in \chi \\ \|f_{0_{\text{res}}}\| &\leq L_{f_{0\,\text{res}}} \|\zeta\| \quad \forall \zeta \in \chi\end{aligned}$$

respectively, where $L_{Y_{\text{res}}}$ and $L_{f_{0\,\text{res}}}$ are positive constants.

[8]See [18, Section 4.4] or [25, Remark 3.6] for details of the projection operator.

To facilitate the subsequent stability analysis, consider the candidate Lyapunov function $V_L : \mathbb{R}^{2n} \times \mathbb{R}^l \times \mathbb{R}^l \times \mathbb{R}^p \to [0, \infty)$ given as

$$V_L(Z) = V(\zeta) + \frac{1}{2}\tilde{W}_c^T \Gamma^{-1} \tilde{W}_c + \frac{1}{2}\tilde{W}_a^T \tilde{W}_a + V_P(Z_P)$$

where $Z \triangleq \left[\zeta^T \ \tilde{W}_c^T \ \tilde{W}_a^T \ Z_P^T\right]^T \in \chi \cup \mathbb{R}^l \times \mathbb{R}^l \times \mathbb{R}^p$. Since the value function V in (17) is positive definite, V_L can be bounded by

$$\underline{\upsilon_L}(\|Z\|) \leq V_L(Z) \leq \overline{\upsilon_L}(\|Z\|) \tag{34}$$

using [26, Lemma 4.3] and (8), where $\underline{\upsilon_L}$, $\overline{\upsilon_L} : [0, \infty) \to [0, \infty)$ are class $\mathcal{K}$ functions. Let $\beta \subset \chi \cup \mathbb{R}^l \times \mathbb{R}^l \times \mathbb{R}^p$ be a compact set.

Theorem 1: Provided Assumptions 1–3 are satisfied, the gains are selected sufficiently large,[9] and the positive constant $K \in \mathbb{R}$ satisfies

$$K < \underline{\upsilon_L}^{-1}\left(\overline{\upsilon_L}(r)\right) \tag{35}$$

where $r \in \mathbb{R}$ is the radius of the compact set β, then the policy in (24) with the NN update laws in (28)–(30) guarantee UUB regulation of the state ζ and UUB convergence of the approximated policies $\hat{u}$ to the optimal policy u^*.

Proof: The time derivative of the candidate Lyapunov function is

$$\begin{aligned}\dot{V}_L = & \frac{\partial V}{\partial \zeta}(Y\theta + f_0) + \frac{\partial V}{\partial \zeta} g(\hat{u} + \hat{\tau}_c) - \tilde{W}_c^T \Gamma^{-1} \dot{\hat{W}}_c \\ & - \frac{1}{2}\tilde{W}_c^T \Gamma^{-1} \dot{\Gamma} \Gamma^{-1} \tilde{W}_c - \tilde{W}_a^T \dot{\hat{W}}_a + \dot{V}_P.\end{aligned} \tag{36}$$

Using (19), $\frac{\partial V}{\partial \zeta}(Y\theta + f_0) = -\frac{\partial V}{\partial \zeta} g(u^* + \tau_c) - r(\zeta, u^*)$. Then,

$$\begin{aligned}\dot{V}_L = & \frac{\partial V}{\partial \zeta} g(\hat{u} + \hat{\tau}_c) - \frac{\partial V}{\partial \zeta} g(u^* + \tau_c) - r(\zeta, u^*) \\ & - \tilde{W}_c^T \Gamma^{-1} \dot{\hat{W}}_c - \frac{1}{2}\tilde{W}_c^T \Gamma^{-1} \dot{\Gamma} \Gamma^{-1} \tilde{W}_c - \tilde{W}_a^T \dot{\hat{W}}_a + \dot{V}_P.\end{aligned}$$

Substituting (28) and (30) for $\dot{\hat{W}}_c$ and $\dot{\hat{W}}_a$, respectively, yields

$$\begin{aligned}\dot{V}_L = & -\zeta^T Q \zeta - u^{*T} R u^* + \frac{\partial V}{\partial \zeta} g \tilde{\tau}_c + \frac{\partial V}{\partial \zeta} g \hat{u} - \frac{\partial V}{\partial \zeta} g u^* \\ & + \tilde{W}_c^T \left[k_{c1} \frac{\omega}{\rho}\delta + \frac{k_{c2}}{N}\sum_{j=1}^{N} \frac{\omega_k}{\rho_k}\delta_k\right] + \tilde{W}_a^T k_a \left(\hat{W}_a - \hat{W}_c\right) \\ & - \frac{1}{2}\tilde{W}_c^T \Gamma^{-1} \left[\left(\beta\Gamma - k_{c1}\Gamma \frac{\omega \omega^T}{\rho}\Gamma\right)\mathbf{1}_{\|\Gamma\| \leq \overline{\Gamma}}\right]\Gamma^{-1}\tilde{W}_c + \dot{V}_P.\end{aligned}$$

Using Young's inequality, (21), (22), (24), (32), and (33) the Lyapunov derivative can be upper bounded as

$$\begin{aligned}\dot{V}_L \leq & -\varphi_\zeta \|\zeta\|^2 - \varphi_c \left\|\tilde{W}_c\right\|^2 - \varphi_a \left\|\tilde{W}_a\right\|^2 - \varphi_\theta \left\|\tilde{\theta}\right\|^2 \\ & - \underline{k_\zeta} \left\|\tilde{\zeta}\right\|^2 + \kappa_a \left\|\tilde{W}_a\right\| + \kappa_c \left\|\tilde{W}_c\right\| + \kappa_\theta \left\|\tilde{\theta}\right\| + \kappa.\end{aligned}$$

[9]For specific details, see Appendix B.

Completing the squares, the upper bound on the Lyapunov derivative may be written as

$$\dot{V}_L \leq -\frac{\varphi_\zeta}{2}\|\zeta\|^2 - \frac{\varphi_c}{2}\left\|\tilde{W}_c\right\|^2 - \frac{\varphi_a}{2}\left\|\tilde{W}_a\right\|^2 - \frac{\varphi_\theta}{2}\left\|\tilde{\theta}\right\|^2 - \underline{k_\zeta}\left\|\tilde{\zeta}\right\|^2 + \frac{\kappa_c^2}{2\varphi_c} + \frac{\kappa_a^2}{2\varphi_a} + \frac{\kappa_\theta^2}{2\varphi_\theta} + \kappa$$

which can be further upper bounded as

$$\dot{V}_L \leq -\alpha\|Z\| \quad \forall \|Z\| \geq K > 0. \tag{37}$$

Using (34), (35), and (37), [26, Th. 4.18] is invoked to conclude that Z is UUB, in the sense that $\limsup_{t\to\infty}\|Z(t)\| \leq \underline{\upsilon_L}^{-1}(\overline{\upsilon_L}(K))$.

Based on the definition of Z and the inequalities in (34) and (37), $\zeta, \tilde{W}_c, \tilde{W}_a \in \mathcal{L}_\infty$. Using the fact that W is upper bounded by a bounded constant and the definition of the NN weight estimation errors, $\hat{W}_c, \hat{W}_a \in \mathcal{L}_\infty$. Using the policy update laws in (30), $\dot{\hat{W}}_a \in \mathcal{L}_\infty$. Since $\hat{W}_c, \hat{W}_a, \zeta \in \mathcal{L}_\infty$ and $\sigma, \nabla\sigma$ are continuous functions of ζ, it follows that $\hat{V}, \hat{u} \in \mathcal{L}_\infty$. From the dynamics in (12), $\dot{\zeta} \in \mathcal{L}_\infty$. By the definition in (26), $\delta \in \mathcal{L}_\infty$. By the definition of the normalized value function update law in (28), $\dot{\hat{W}}_c \in \mathcal{L}_\infty$. ■

VII. Experimental Validation

Validation of the proposed controller is demonstrated with experiments conducted at Ginnie Springs in High Springs, FL, USA. Ginnie Springs is a second-magnitude spring discharging 142 million liters of freshwater daily with a spring pool measuring 27.4 m in diameter and 3.7 m deep [27]. Ginnie Springs was selected to validate the proposed controller because of its relatively high flow rate and clear waters for vehicle observation. For clarity of exposition[10] and to remain within the vehicle' s depth limitations,[11] the developed method is implemented on an AUV, where the surge, sway, and yaw are controlled by the algorithm represented in (31).

A. Experimental Platform

Experiments were conducted on an AUV, SubjuGator 7, developed at the University of Florida. The AUV, shown in Fig. 1, is a small two man portable AUV with a mass of 40.8 kg. The vehicle is overactuated with eight bidirectional thrusters. The vehicle includes a 2.13-GHz server grade quad-core processor. The suite of navigation sensors include an inertial measurement unit, a DVL, a depth sensor, and a digital compass. The navigation vessel also includes an embedded 720-MHz processor for preprocessing and packaging navigation data. The vehicle' s software runs within the Robot Operating System framework in the central pressure vessel. For the experiment, three main software nodes were used: navigation, control, and thruster mapping nodes. The navigation node receives packaged navigation data from the navigation pressure vessel where an extended Kalman filter estimates the vehicle' s full state at 50 Hz. The controller node contains the developed controller and system identifier. The desired force and moment produced by the controller are mapped to the eight thrusters using a least-squares minimization algorithm in the thruster mapping node. Further details regarding the vehicle construction are given in [5].

Fig. 1. SubjuGator 7 AUV operating at Ginnie Springs, FL, USA.

B. Controller Implementation

The implementation of the developed method involves: system identification, value function iteration, and control iteration. Implementing the system identifier requires (4), (6), and the dataset described in Assumption 2. The dataset in Assumption 2 was collected in a swimming pool. The vehicle was commanded to track an exciting trajectory with a RISE controller [5] while the state-action pairs were recorded. The recorded data were trimmed to a subset of 40 sampled points that were selected to maximize the minimum singular value of $\begin{bmatrix} Y_1 & Y_2 & \dots & Y_j \end{bmatrix}$ as in [15, Algorithm 1].

Evaluating the extrapolated Bellman error in (26) with each control iteration is computational expensive. Due to the limited computational resources available on-board the AUV, the value function weights were updated at a slower rate (i.e., 5 Hz) than the main control loop (implemented at 50 Hz). The developed controller was used to control the surge, sway, and yaw states of the AUV, and a nominal controller was used to regulate the remaining states. Assumption 1 was not applicable in the experiments since a nominal controller was used to regulate heave, roll, and pitch. Assumption 2 is valid for the experiments because a full rank history stack is developed prior to beginning the experiment, and a singular value maximization algorithm is used to ensure it remains positive definite. Assumption 3 was validated during run time for each control iteration during the experiment.

The vehicle uses water profiling data from the DVL to measure the relative water velocity near the vehicle in addition to bottom tracking data for the state estimator. By using the state estimator, water profiling data, and recorded data, the

[10]The number of basis functions and weights required to support a six DOF model greatly increases from the set required for the three DOF model. The increased number of parameters and complexity reduces the clarity of this proof of principal experiment.

[11]The vehicle's DVL has a minimum height over bottom of approximately 3 m that is required to measure water velocity. A minimum depth of approximately 0.5 m is required to remove the vehicle from surface effects. With the depth of the spring nominally 3.7 m, a narrow window of about 20 cm is left operate the vehicle in heave.

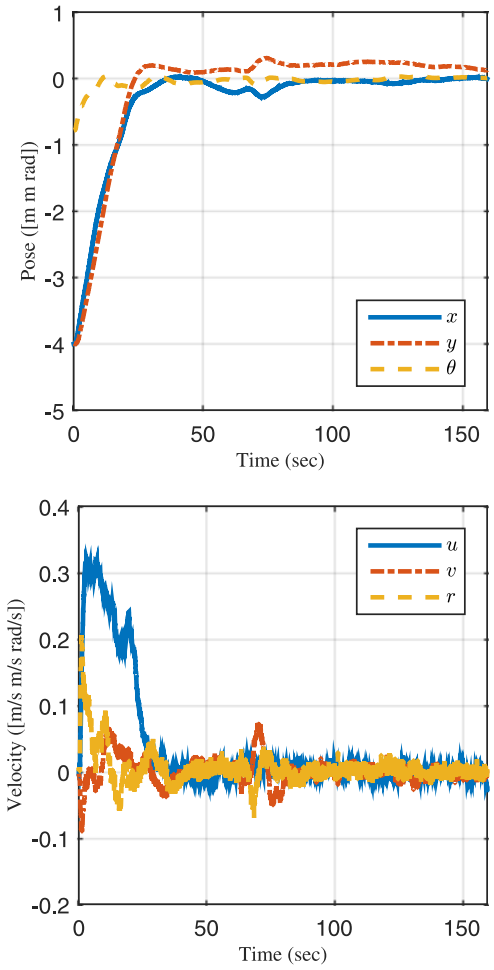


Fig. 2. Inertial position error η (top) and body-fixed, over ground, velocity error ν (bottom) of the AUV.

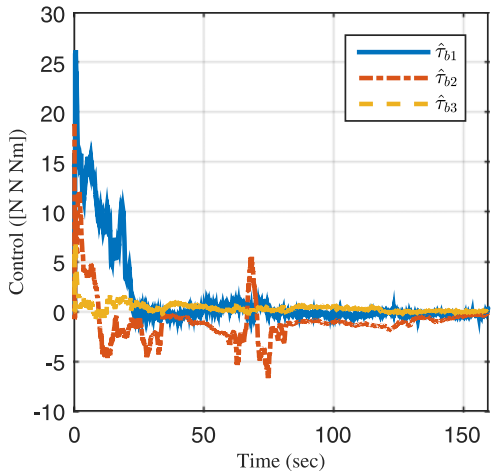


Fig. 3. Body-fixed total control effort $\hat{\tau}_b$ commanded about the center of mass of the vehicle.

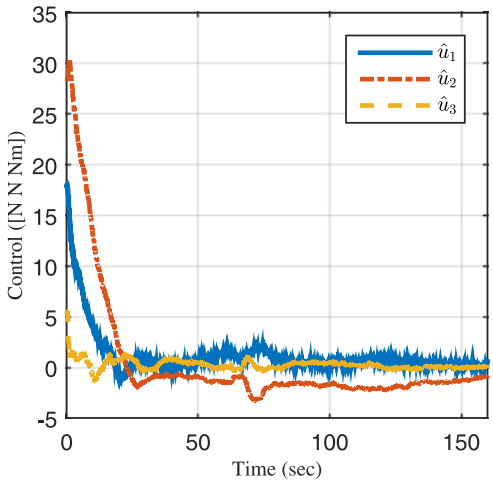


Fig. 4. Body-fixed optimal control effort $\hat{u}$ commanded about the center of mass of the vehicle.

equations used to implement the proposed controller, i.e., (4), (6), (24), (26), and (28)–(31), only contain known or measurable quantities.

C. Results

The vehicle was commanded to hold a station near the vent of Ginnie Spring. An initial condition of $\zeta(t_0) = [4\,\text{m}\ \ 4\,\text{m}\ \ \frac{\pi}{4}\,\text{rad}\ \ 0\,\text{m/s}\ \ 0\,\text{m/s}\ \ 0\,\text{rad/s}]^T$ was given to demonstrate the method's ability to regulate the state. The optimal control weighting matrices were selected to be $Q = \text{diag}([20, 50, 20, 10, 10, 10])$ and $R = I_{3\times 3}$. The uncertain parameters $\theta = \begin{bmatrix} X_u & Y_v & Y_r & N_v & N_r & X_{\dot{u}} & Y_{\dot{v}} & Y_{\dot{r}} \end{bmatrix}$ in the C_A and D_A matrices are as defined in [16, Sec. 7.5]. The system identifier adaptation gains were selected to be $k_\zeta = 25 \times I_{6\times 6}$, $k_\theta = 12.5$, and $\Gamma_\theta = \text{diag}([187.5, 937.5, 37.5, 37.5, 37.5, 37.5, 37.5, 37.5])$. The parameter estimate was initialized with $\hat{\theta}(t_0) = 0_{8\times 1}$. The NN weights were initialized to match the ideal values for the linearized optimal control problem, which is obtained by solving the algebraic Riccati equation with the dynamics linearized about the station. The policy adaptation gains were chosen to be $k_{c1} = 0.25$, $k_{c2} = 0.5$, $k_a = 1$, $k_p = 0.25$, and $\beta = 0.025$. The adaptation matrix was initialized to $\Gamma_0 = 400 \times I_{21\times 21}$. The Bellman error was extrapolated to sampled states that were uniformly selected throughout the state space in the vehicle' s operating domain.

Fig. 2 illustrates the ability of the generated policy to regulate the state in the presence of the spring's current. Fig. 3 illustrates the total control effort applied to the body of the vehicle, which includes the estimate of the current compensation term and approximate optimal control. Fig. 4 illustrates the output of the approximate optimal policy for the residual system. Fig. 5 illustrates the convergence of the parameters of the system identifier and Fig. 6 illustrates convergence of the NN weights representing the value function.

The anomaly seen at ~70 s in the total control effort (Fig. 3) is attributed to a series of incorrect current velocity measurements. The corruption of the current velocity measurements is possibly due in part to the extremely low turbidity in the spring and/or relatively shallow operating depth. Despite presence of unreliable current velocity measurements the vehicle was able to regulate the vehicle to its station. The results demonstrate the developed method's ability to concurrently identify the unknown hydrodynamic parameters and generate an approximate

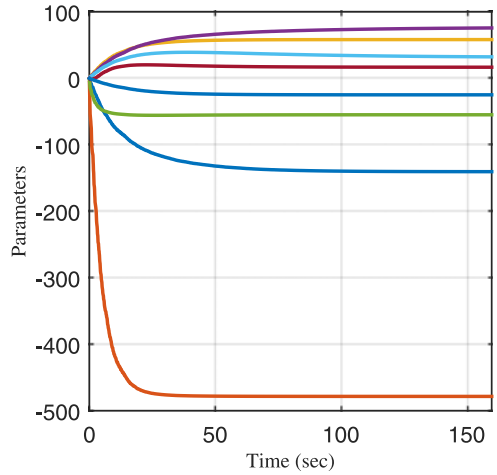


Fig. 5. Identified system parameters determined for the vehicle online. The parameter definitions may be found in [16, Example 6.2 and Equation 6.1].

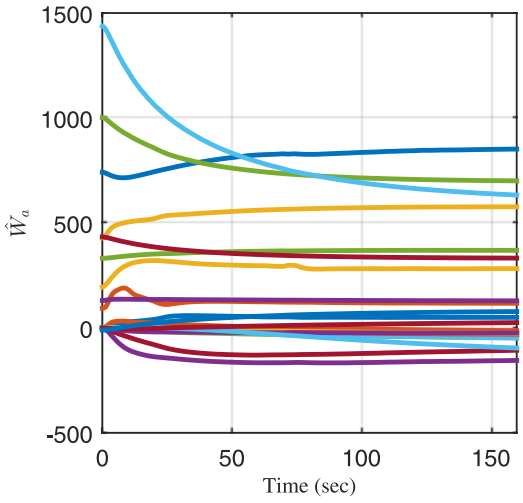


Fig. 6. Actor NN weight estimates, $\hat{W}_a$.

optimal policy using the identified model. The vehicle follows the generated policy to achieve its station-keeping objective using industry standard navigation and environmental sensors (i.e., IMU, DVL).

VIII. Conclusion

The online approximation of an optimal control strategy is developed to enable station keeping by an AUV. The solution to the HJB equation is approximated using ADP. The hydrodynamic effects are identified online with a concurrent learning-based system identifier. Leveraging the identified model, the developed strategy simulates exploration of the state space to learn the optimal policy without the need of a persistently exciting trajectory. A Lyapunov-based stability analysis concludes UUB convergence of the states and UUB convergence of the approximated policies to the optimal polices. Experiments in a central Florida second-magnitude spring demonstrate the ability of the controller to generate and execute an approximate optimal policy in the presence of a time-varying irrational current.

Appendix A
Extension to Constant Earth-Fixed Current

In the case where the earth-fixed current is constant, the effects of the current may be included in the development of the optimal control problem. The body-relative current velocity $\nu_c(\zeta)$ is state dependent and may be determined from

$$\dot{\eta}_c = \begin{bmatrix} \cos(\psi) & -\sin(\psi) \\ \sin(\psi) & \cos(\psi) \end{bmatrix} \nu_c$$

where $\dot{\eta}_c \in \mathbb{R}^n$ is the known constant current velocity in the inertial frame. The functions $Y_{\text{res}}\theta$ and $f_{0_{\text{res}}}$ in (11) can then be redefined as

$$Y_{\text{res}}\theta \triangleq \begin{bmatrix} 0 \\ -M^{-1}C_A(-\nu_c)\nu_c - M^{-1}D_A(-\nu_c)\nu_c \dots \\ -M^{-1}C_A(\nu_r)\nu_r - M^{-1}D_A(\nu_r)\nu_r \end{bmatrix}$$

$$f_{0_{\text{res}}} \triangleq \begin{bmatrix} J_E\nu \\ -M^{-1}C_{RB}(\nu)\nu - M^{-1}G(\eta) \end{bmatrix}$$

respectively. The control vector u is

$$u = \tau_b - \tau_c$$

where $\tau_c(\zeta) \in \mathbb{R}^n$ is the control effort required to keep the vehicle on station given the current and is redefined as

$$\tau_c \triangleq -M_A\dot{\nu}_c - C_A(-\nu_c)\nu_c - D_A(-\nu_c)\nu_c.$$

Appendix B
Stability Analysis Terms

To facilitate the development of the sufficient gain conditions, the following terms are defined:

$$\varphi_\zeta = \underline{q} - \frac{k_{c1}\sup_{Z\in\beta}\|\epsilon'\|\left(L_{Y_{\text{res}}}\|\theta\| + L_{f_{0_{\text{res}}}}\right)}{2} - \frac{L_{Y_c}\|g\|\left(\|W\|\sup_{Z\in\beta}\|\sigma'\| + \sup_{Z\in\beta}\|\epsilon'\|\right)}{2}$$

$$\varphi_c = \frac{k_{c2}}{N}\underline{c} - \frac{k_a}{2} - \frac{k_{c1}\sup_{Z\in\beta}\|\epsilon'\|\left(L_{Y_{\text{res}}}\|\theta\| + L_{f_{0_{\text{res}}}}\right)}{2} - \frac{k_{c1}L_Y\sup_{Z\in\beta}\|\zeta\|\sup_{Z\in\beta}\|\sigma'\|\|W\|}{2} - \frac{\frac{k_{c2}}{N}\sum_{j=1}^{n}\left(\left\|Y_{\text{res}_j}\sigma'_j\right\|\right)\|W\|}{2}$$

$$\varphi_a = \frac{k_a}{2}$$

$$\varphi_\theta = k_\theta \underline{y} - \frac{\frac{k_{c2}}{N}\sum_{k=1}^{N}\left(\left\|Y_{\text{res}_k}\sigma'_k\right\|\right)\|W\|}{2} - \frac{L_{Y_c}\|g\|\left(\|W\|\sup_{Z\in\beta}\|\sigma'\| + \sup_{Z\in\beta}\|\epsilon'\|\right)}{2} - \frac{k_{c1}L_{Y_{\text{res}}}\|W\|\sup_{Z\in\beta}\|\zeta\|\sup_{Z\in\beta}\|\sigma'\|}{2}$$

$$\kappa_c = \sup_{Z\in\beta}\left\|\frac{k_{c2}}{4N}\sum_{j=1}^{N}\tilde{W}_a^T G_{\sigma_j}\tilde{W}_a + \frac{k_{c1}}{4}\tilde{W}_a^T G_\sigma \tilde{W}_a + k_{c1}\epsilon' G\sigma'^T W + \frac{k_{c1}}{4}\epsilon' G\epsilon'^T + \frac{k_{c2}}{N}\sum_{k=1}^{N}E_k\right\|$$

$$\kappa_a = \sup_{Z\in\beta}\left\|\frac{1}{2}W^T G_\sigma + \frac{1}{2}\epsilon' G\sigma'^T\right\|$$

$$\kappa_\theta = k_\theta d_\theta$$

$$\kappa = \sup_{Z\in\beta}\left\|\frac{1}{4}\epsilon' G\epsilon'^T\right\|.$$

When Assumptions 2 and 3 and the sufficient gain conditions

$$\underline{q} > \frac{k_{c1}\sup_{Z\in\beta}\|\epsilon'\|\left(L_{Y_{\text{res}}}\|\theta\| + L_{f_{0_{\text{res}}}}\right)}{2} + \frac{L_{Y_c}\|g\|\left(\|W\|\sup_{Z\in\beta}\|\sigma'\| + \sup_{Z\in\beta}\|\epsilon'\|\right)}{2}$$

$$\underline{c} > \frac{N}{k_{c2}}\left(\frac{k_{c1}\sup_{Z\in\beta}\|\epsilon'\|\left(L_{Y_{\text{res}}}\|\theta\| + L_{f_{0_{\text{res}}}}\right)}{2} + \frac{k_a}{2} + \frac{k_{c1}L_Y\sup_{Z\in\beta}\|\zeta\|\sup_{Z\in\beta}\|\sigma'\|\|W\|}{2} + \frac{\frac{k_{c2}}{N}\sum_{k=1}^{N}\left(\left\|Y_{\text{res}_k}\sigma'_k\right\|\right)\|W\|}{2}\right)$$

$$\underline{y} > \frac{1}{k_\theta}\left(\frac{\frac{k_{c2}}{N}\sum_{k=1}^{N}\left(\left\|Y_{\text{res}_k}\sigma'_k\right\|\right)\|W\|}{2} + \frac{L_{Y_c}\|g\|\left(\|W\|\sup_{Z\in\beta}\|\sigma'\| + \sup_{Z\in\beta}\|\epsilon'\|\right)}{2} + \frac{k_{c1}L_{Y_{\text{res}}}\|W\|\sup_{Z\in\beta}\|\zeta\|\sup_{Z\in\beta}\|\sigma'\|}{2}\right)$$

are satisfied, the constant K defined as

$$K \triangleq \sqrt{\frac{\kappa_c^2}{2\alpha\varphi_c} + \frac{\kappa_a^2}{2\alpha\varphi_a} + \frac{\kappa_\theta^2}{2\alpha\varphi_\theta} + \frac{\kappa}{\alpha}}$$

is positive, where $\alpha \triangleq \frac{1}{2}\min\left\{\varphi_\zeta, \varphi_c, \varphi_a, \varphi_\theta, 2\underline{k_\zeta}\right\}$.

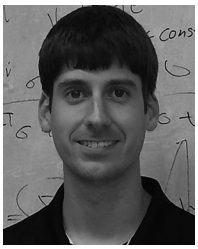
ACKNOWLEDGMENT

The authors would like to thank Ginnie Springs Outdoors, LLC, who provided access to Ginnie Springs for the validation of the developed controller.

REFERENCES

[1] A. J. Sorensen, "A survey of dynamic positioning control systems," *Annu. Rev. Control*, vol. 35, pp. 123–136, 2011.
[2] T. Fossen and A. Grovlen, "Nonlinear output feedback control of dynamically positioned ships using vectorial observer backstepping," *IEEE Trans. Control Systems Technol.*, vol. 6, no. 1, pp. 121–128, Jan. 1998.
[3] E. Sebastian and M. A. Sotelo, "Adaptive fuzzy sliding mode controller for the kinematic variables of an underwater vehicle," *J. Intell. Robot. Syst.*, vol. 49, no. 2, pp. 189–215, 2007.
[4] E. Tannuri, A. Agostinho, H. Morishita, and L. Moratelli Jr, "Dynamic positioning systems: An experimental analysis of sliding mode control," *Control Eng. Pract.*, vol. 18, pp. 1121–1132, 2010.
[5] N. Fischer, D. Hughes, P. Walters, E. Schwartz, and W. E. Dixon, "Nonlinear RISE-based control of an autonomous underwater vehicle," *IEEE Trans. Robot.*, vol. 30, no. 4, pp. 845–852, Aug. 2014.
[6] R. W. Beard and T. W. Mclain, "Successive galerkin approximation algorithms for nonlinear optimal and robust control." *Int. J. Control*, vol. 71, pp. 717–743, 1998.
[7] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for approximate optimal regulation," *Automatica*, vol. 64, pp. 94–104, 2016.
[8] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 89–92, Jan. 2013.
[9] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Netw.*, vol. 22, no. 3, pp. 237–246, 2009.
[10] K. Vamvoudakis and F. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.
[11] H. Modares, F. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, Oct. 2013.
[12] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.
[13] M. Johnson, S. Bhasin, and W. E. Dixon, "Nonlinear two-player zero-sum game approximate solution using a policy iteration algorithm," in *Proc. IEEE Conf. Decis. Control*, 2011, pp. 142–147.
[14] K. Vamvoudakis and F. Lewis, "Online neural network solution of nonlinear two-player zero-sum games using synchronous policy iteration," in *Proc. 49th IEEE Conf. Decis. Control*, 2010, pp. 3040–3047.
[15] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *Int. J. Adapt. Control Signal Process.*, vol. 27, no. 4, pp. 280–301, 2013.
[16] T. I. Fossen, *Handbook of Marine Craft Hydrodynamics and Motion Control*. Hoboken, NJ, USA: Wiley, 2011.
[17] S. Sastry and A. Isidori, "Adaptive control of linearizable systems," *IEEE Trans. Autom. Control*, vol. 34, no. 11, pp. 1123–1131, Nov. 1989.
[18] P. Ioannou and J. Sun, *Robust Adaptive Control*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1996.
[19] G. V. Chowdhary and E. N. Johnson, "Theory and flight-test validation of a concurrent-learning adaptive controller," *J. Guid. Control Dyn.*, vol. 34, no. 2, pp. 592–607, Mar. 2011.
[20] R. Kamalapurkar, "Model-based reinforcement learning for online approximate optimal control," Ph.D. dissertation, University of Florida, Gainesville, FL, USA, 2014.
[21] A. Gelb, *Applied Optimal Estimation*. Cambridge, MA, USA: The MIT press, 1974.
[22] D. Kirk, *Optimal Control Theory: An Introduction*. Mineola, NY, USA: Dover, 2004.
[23] P. M. Patre, W. MacKunis, K. Kaiser, and W. E. Dixon, "Asymptotic tracking for uncertain dynamic systems via a multilayer neural network feedforward and RISE feedback control structure," *IEEE Trans. Autom. Control*, vol. 53, no. 9, pp. 2180–2185, Oct. 2008.
[24] R. Kamalapurkar, J. Klotz, and W. E. Dixon, "Concurrent learning-based online approximate feedback Nash equilibrium solution of N-player nonzero-sum differential games," *IEEE/CAA J. Autom. Sin.*, vol. 1, no. 3, pp. 239–247, Jul. 2014.

[25] W. E. Dixon, A. Behal, D. M. Dawson, and S. Nagarkatti, *Nonlinear Control of Engineering Systems: A Lyapunov-Based Approach*. Boston, MA, USA: Birkhauser, 2003.
[26] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
[27] W. Schmidt, "Springs of Florida," Florida Geological Survey, Bulletin 66, 2004.



Patrick Walters received the Ph.D. degree in mechanical engineering from the University of Florida, Gainesville, FL, USA, in 2015.

He is currently a staff researcher at the Naval Surface Warfare Center Panama City Division, Panama City, FL, USA. His research interests include reinforcement learning-based feedback control, approximate dynamic programming, and robust control of uncertain nonlinear systems with a focus on the application to underwater vehicles.

Rushikesh Kamalapurkar (SM'17) received the M.S. and Ph.D. degrees in mechanical engineering from the University of Florida, Gainesville, FL, USA, in 2011 and 2014, respectively.

After working for a year as a postdoctoral researcher with Dr. Warren E. Dixon, he was appointed as the 2015–2016 MAE postdoctoral teaching fellow. In 2016 he joined the School of Mechanical and Aerospace Engineering at the Oklahoma State University as an Assistant professor. His primary research interest is intelligent, learning-based optimal control of uncertain nonlinear dynamical systems. He has published 3 book chapters, 19 peer reviewed journal papers and 21 peer reviewed conference papers.

Dr. Kamalapurkar's work has been recognized by the 2015 University of Florida Department of Mechanical and Aerospace Engineering Best Dissertation Award, and the 2014 University of Florida Department of Mechanical and Aerospace Engineering Outstanding Graduate Research Award.

Forrest Voight received the M.S. degree in electrical engineering from University of Florida, Gainesville, FL, USA, in 2016.

He is currently the CEO of Sylphase LLC, Gainesville, FL, USA, where his primary role includes development and testing of satellite-aided inertial navigation systems. His research interests include state estimation of nonlinear systems, especially in the context of satellite and inertial navigation, and inertial sensor design.

Eric M. Schwartz (S'82–M'85–SM'02) received the B.S. degrees in both electrical and mechanical engineering, the M.S. in electrical engineering, and the Ph.D. degree in electrical and computer engineering, all from University of Florida, Gainesville, FL, USA, in 1984, 1989, and 1995, respectively.

During three semesters in 1980 and 1981, he was with Bendix Avionics in Fort Lauderdale, FL, USA. In the summer of 1982, he was with Universal Security Instruments in Baltimore, MD, USA. In the summer of 1983, he was with International Business Machines in Boca Raton, FL, USA. In the summer of 1987, he was with SASIAM and the Advanced Robotics Research Group of Technopolis, Bari, Italy. In the summer of 1988, he was with Allen-Bradley Corporation, Cleveland, OH, USA. In 1989–1993, he was with the Electronic Communications Laboratory, University of Florida. In 1985, he was with the Machine Intelligence Laboratory (MIL), University of Florida, as a Graduate Student Researcher and became a Postgraduate Researcher for MIL in 1995, the Assistant Director of MIL in 1998, the Associate Director in 2004, and the Director in 2016. In 1995–1997, he was a Visiting Assistant Professor in Electrical and Computer Engineering, University of Florida, was a Lecturer from 1997 to 2005, was a Senior Lecturer from 2005 to 2008, and became a Master Lecturer in 2008.

Dr. Schwartz has been Treasurer or Secretary of the IEEE Gainesville Section since 2002 and the Faculty Advisor of the IEEE Gainesville Section Student Branch since 2002. He received the 2002–2003 University of Florida Teacher of the Year Award. He has directed many robot teams, including five world champions (three RoboSub, one RoboBoat, and one Maritime RobotX Challenge in 2016). In 14 of the 25 world championships in which his teams have competed, they earned third place or better.

Warren E. Dixon (M'94–F'16) received the Ph.D. degree in electrical engineering from Clemson University, Clemson, SC, USA, in 2000.

He was a Research Staff Member and Eugene P. Wigner Fellow at Oak Ridge National Laboratory, Oak Ridge, TN, USA, until 2004, when he joined the Mechanical and Aerospace Engineering Department, University of Florida, Gainesville, FL, USA. His main research interest has been the development and application of Lyapunov-based control techniques for uncertain nonlinear systems.

Dr. Dixon's work has been recognized by the 2009 and 2015 American Automatic Control Council O. Hugo Schuck (Best Paper) Award, the 2013 Fred Ellersick Award for Best Overall MILCOM Paper, a 2012–2013 University of Florida College of Engineering Doctoral Dissertation Mentoring Award, the 2011 American Society of Mechanical Engineers Dynamics Systems and Control Division Outstanding Young Investigator Award, the 2006 IEEE Robotics and Automation Society Early Academic Career Award, an NSF CAREER Award, the 2004 Department of Energy Outstanding Mentor Award, and the 2001 ORNL Early Career Award for Engineering Achievement. He is a Fellow of ASME, an IEEE Control Systems Society Distinguished Lecturer, and served as the Director of Operations for the Executive Committee of the IEEE CSS Board of Governors (2012–2015). He was awarded the Air Force Commander's Public Service Award in 2016 for his contributions to the U.S. Air Force Science Advisory Board. He is currently or formerly an Associate Editor for *ASME Journal of Journal of Dynamic Systems, Measurement and Control, Automatica*, IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS: PART B CYBERNETICS, and *International Journal of Robust and Nonlinear Control.*