Control of a Realistic Wave Energy Converter Model Using Least-Squares Policy Iteration

Enrico Anderlini, David I. M. Forehand, Elva Bannon, and Mohammad Abusara

***Abstract*—An algorithm has been developed for the resistive control of a nonlinear model of a wave energy converter using least-squares policy iteration, which incorporates function approximation, with tabular and radial basis functions being used as features. With this method, the controller learns the optimal power take-off damping coefficient in each sea state for the maximization of the mean generated power. The performance of the algorithm is assessed against two online reinforcement learning schemes: Q-learning and SARSA. In both regular and irregular waves, least-squares policy iteration outperforms the other strategies, especially when starting from unfavorable conditions for learning. Similar performance is observed for both basis functions, with a smaller number of radial basis functions underfitting the Q-function. The shorter learning time is fundamental for a practical application on a real wave energy converter. Furthermore, this paper shows that least-squares policy iteration is able to maximize the energy absorption of a wave energy converter despite strongly nonlinear effects due to its model-free nature, which removes the influence of modeling errors. Additionally, the floater geometry has been changed during a simulation to show that reinforcement learning control is able to adapt to variations in the system dynamics.**

***Index Terms*—Function approximation, radial basis function (RBF), reinforcement learning (RL), resistive control, wave energy converter (WEC).**

I. Introduction

WAVE energy has the potential to become a significant contributor to the future energy mix thanks to a resource of up to 2.1 TW of power worldwide [1], with a consequent reduction in greenhouse gas emissions. Nevertheless, wave energy converters (WECs) are not economically viable yet, despite numerous designs having been proposed over the years. A review of some of the most promising, recent technologies can be found in [2]. The design of an effective control scheme can considerably reduce the levelised cost of energy associated with WECs, since it can bring about a gain in energy absorption with little additional hardware costs.

Since the 1970s, multiple control strategies have been studied so as to maximise energy extraction. Thorough reviews of the topic can be found in [3] for the initial analyses and in [4] for more recent developments. By achieving resonance between the device and the incident waves, complex-conjugate control would result theoretically in optimal power absorption [3]. However, this is infeasible in practice because of the resulting large motions of the WEC in energetic sea states and the associated high loads. Hence, alternative control algorithms have been developed, which limit the motions, forces and power ratings of the device [4]. It is possible to differentiate between two main types of control schemes: time-averaged and real-time.

Time-averaged strategies assume stationary wave conditions over a prescribed time, over which a constant, optimal control setting is used [5]. In the case of reactive control, this is represented by the combination of damping and stiffness coefficients of the Power Take-Off (PTO) unit that maximise energy generation in each sea state. A specific case occurs for zero stiffness: resistive control. The optimal values are found through preliminary simulations, which can constrain the force and displacement, and then stored in a look-up table. Whereas time-averaged schemes may be less efficient than real-time strategies, their computational cost is lower.

Real-time algorithms consist of applying an optimal control action at every time instant such that it is expected to maximize energy generation over a short (in the order of one wave cycle) future time horizon [4]. Examples are latching [6], declutching [7], simple-but-effective [8], and model predictive control [9]–[11]. Whereas simple-but-effective control is computationally light, since it relies on classical closed-loop controllers, non-linear model predictive control and latching and declutching control based on Pontryagin's principle can present a high computational cost associated with their real-time optimization. Solutions to reduce the computational costs are to use a linear model in model predictive control and a moving window in latching and declutching control, as for instance proposed by [12]. Furthermore, the control operation is strongly affected by the accuracy of the wave excitation force forecast.

As the performance of all aforementioned control schemes depends on the quality of the model of the device dynamics, modelling errors can decrease the generated power as well as cause damage to the machines if the physical limits are exceeded

Manuscript received November 25, 2016; revised March 22, 2017; accepted April 17, 2017. Date of publication April 25, 2017; date of current version September 15, 2017. This work was supported in part by the Energy Technologies Institute and the Research Councils Energy Programme under Grant EP/J500847/, in part by the Engineering and Physical Sciences Research Council under Grant EP/J500847/1, and in part by Wave Energy Scotland. Paper no. TSTE-00924-2016. *(Corresponding author: Enrico Anderlini.)*

E. Anderlini is with the Industrial Doctoral Centre, Offshore Renewable Energy, Edinburgh EH9 3JL, U.K. (e-mail: E.Anderlini@ed.ac.uk).

D. I. M. Forehand is with the Institute of Energy Systems, University of Edinburgh, Edinburgh EH9 3DW, U.K. (e-mail: D.Forehand@ed.ac.uk).

E. Bannon is with Wave Energy Scotland, Highlands and Islands Enterprise, Inverness IV2 5NA, U.K. (e-mail: elva.bannon@hient.co.uk).

M. Abusara is with the College of Engineering, University of Exeter, Cornwall TR10 9FE, U.K. (e-mail: M.Abusara@exeter.ac.uk).

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/TSTE.2017.2696060

in practice. Whereas a hierarchical robust controller has been used to decrease the sensitivity of simple-but-effective control to modelling errors and non-linear effects [13], the other control strategies are negatively affected by modelling errors. An alternative approach to robust control based on fuzzy logic has been proposed by [14]. Similarly, the authors proposed, in a previous study, the application of an alternative strategy, reinforcement learning (RL), to the control of WECs. This scheme does not rely on a model of the WEC dynamics to obtain the control action and is thus able to adapt to changes in the system response due to ageing, e.g. the build-up of marine biofouling. Although the algorithm described in [15] can be implemented on an actual device, a test-case numerical study was run in that paper using a linear model of a point absorber, a well known WEC technology which comprises of a float whose size is small compared with the characteristic wavelength [2].

In this article, RL is applied to the control of a realistic, non-linear model of a WEC, whose accuracy was validated by the developers against measurements on a prototype device [16]. In the absence of costly experimental measurements, this study enables the assessment of the convergence behaviour of RL when non-linear effects are important. As in [16], only damping, or resistive, control is analysed. This is known to be inferior in performance to fully reactive control, which is in fact treated by most other real-time control schemes, as described in [4]. However, the practical implementation of resistive control is simple. Additionally, the optimization of only the damping coefficient results in a simpler framework to demonstrate the applicability of RL for the control for WECs, which can then be extended to the treatment of combined damping and stiffness control. In particular, here resistive control is implemented using least-squares policy iteration (LSPI), an efficient RL algorithm [17]. Furthermore, its performance is assessed against two simpler RL algorithms, SARSA and Q-learning [18]. In addition, the effectiveness of function approximation in reducing the learning time has been assessed using LSPI with radial basis functions (RBFs) [17].

II. Resistive Control of the Seabased WEC

A. System Description

The Seabased device is a point absorber with a direct-drive PTO system. The development and testing of a number of full-scale prototypes at Uppsala University is well described in the literature [19]–[23]. The version studied in [16] is analysed in this article, although the generator is now connected to the electrical grid.

Fig. 1 shows a diagram of the device, which is inspired by [16]. A small float, excited by incident waves, drives a linear, permanent-magnet generator along vertical rails. The two bodies are connected by a mooring line. When the distance between the float and the translator decreases, the mooring line goes slack and the translator is pulled downwards by a dedicated spring. Additionally, springs at the upper and lower end stops prevent the translator from breaking the casing in large waves. The motion of the magnet induces electrical current in the coils wound around the stator. Power absorption is controlled through

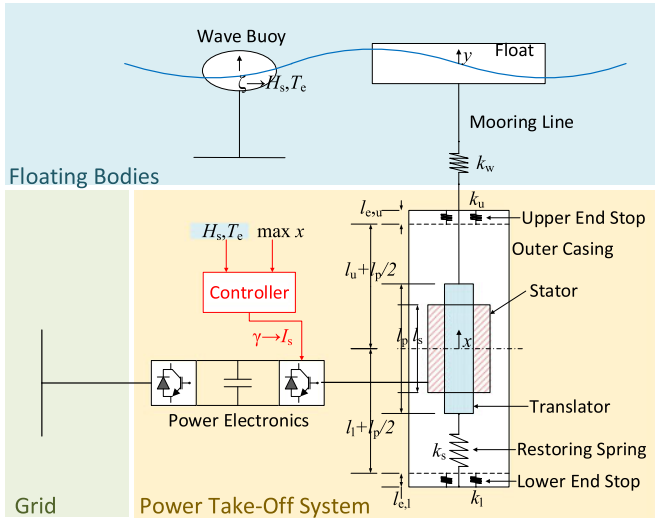

Fig. 1. Diagram of the prototype Seabased WEC.

TABLE I
Distance Between Kernels, Bandwidth and Number of Kernels Used in the Study of LSPI With RBFs

δ_c (kNs/m)	μ (kNs/m)	M
10	10	10
10	20	10
20	10	5
20	20	5
20	40	5

a power electronic converter by setting the stator current I_s to be proportional to the velocity of the translator. A second power electronic converter controls the voltage across the capacitor between the converters by setting the grid current. The wave elevation ζ is measured through a wave buoy sited 80 m from the prototype at the Lysekil wave energy research site [21].

In Fig. 1, the same naming convention as in [16] is held, with the values of the variables quantities being given in Table I in [16]. In addition, $l_{e,u} = 0.25$ m and $l_{e,l} = 0.14$ m are a measure of the end stops length, as given in [21].

B. Mathematical Model

A weakly non-linear mathematical model of the system dynamics has been developed by [16]. Although the float is free to move in all directions in reality, only the heave degree of freedom is analysed because the influence of the other motions is negligible [24]. Defining y and x as the float and translator displacement respectively, the motions of the two bodies are described by the following system of equations

$$(m_b + m_\infty)\,\ddot{y}(t) = F_e(t) - F_r(t) - F_h(t) - F_w(t), \quad (1)$$

$$m_p \ddot{x}(t) = F_w(t) - F_{em}(t) - F_s(t) + F_u(t) + F_l(t), \quad (2)$$

where t indicates time, m_b and m_p the mass of the float and piston respectively, and m_∞ the added mass of the float. F_e is the incident and diffracted wave excitation force, F_r part of the

float radiation force, F_{h} the hydrostatic restoring force, F_{w} the tension in the wire connecting the float to the translator, F_{em} the electromotive force, F_{s} the force of the restoring spring in Fig. 1, F_{u} and F_{l} the spring force of the upper and lower end stops, respectively. The non-linearities are associated with F_{w}, where compression effects are ignored, F_{u} and F_{l}, which are activated only if the end stops are reached, and F_{em}, which depends on the exposure of the translator to the stator.

The electromotive, or control, force F_{em} is discussed in the next section. The hydrostatic force is calculated as

$$F_{\mathrm{h}}(t) = \rho g S_{\mathrm{w}} y(t), \tag{3}$$

where $\rho = 1025\ \mathrm{kg/m^3}$ is the seawater density, $g = 9.81\ \mathrm{m/s^2}$ the gravitational acceleration and S_{w} the float waterplane area.

The F_{r} part of radiation force is approximated through a state-space system so as to reduce its computational cost [25]

$$\dot{\boldsymbol{x}}_{\mathrm{ss}}(t) = \boldsymbol{A}_{\mathrm{ss}}\boldsymbol{x}_{\mathrm{ss}}(t) + \boldsymbol{B}_{\mathrm{ss}}\dot{y}(t) \tag{4}$$

$$F_{\mathrm{r}}(t) \approx \boldsymbol{C}_{\mathrm{ss}}\boldsymbol{x}_{\mathrm{ss}}(t). \tag{5}$$

The matrices $\boldsymbol{A}_{\mathrm{ss}}$, $\boldsymbol{B}_{\mathrm{ss}}$ and $\boldsymbol{C}_{\mathrm{ss}}$ are calculated with frequency-domain system identification, as described in [25]. F_{r} is given by the convolution of the product of the radiation impulse response function and the float velocity [26]. Furthermore, the radiation impedance function has been computed using the commercial software WAMIT for the vertical cylinder geometry described in [16] for circular wave frequency values ranging from 0 rad/s to 10 rad/s in steps of 0.005 rad/s. Furthermore, m_∞ has been calculated for the infinite wave frequency case.

Similarly, the wave excitation force F_{e} is given by the convolution integral of the product of the wave elevation at the float location and the diffraction impulse response function [26], which has also been computed using WAMIT. In irregular waves, a number of wave components have been superimposed to obtain ζ. The amplitude of each wave has been obtained from a wave spectrum, sampled at a circular wave frequency step of 0.005 rad/s in order to prevent repeating the wave trace within a 15-minute window, as the value is less than the Nyquist frequency [27]. In order to obtain longer time series, individual wave traces generated using a different seed to the random number generator have been joined. The connections are smoothed using a 200-point filter over the last and first 20 s of each wave trace.

The force of the spring connected to the translator is expressed as [16]

$$F_{\mathrm{s}} = F_0 + k_{\mathrm{s}}x, \tag{6}$$

where F_0 is a static force due to precharging, and k_{s} is the spring stiffness. Ignoring compression effects, the wire force is given by [16]

$$F_{\mathrm{w}} = \begin{cases} -k_{\mathrm{w}}(y-x) & \text{if } y > x \\ 0 & \text{else} \end{cases}, \tag{7}$$

with k_{w} being the wire stiffness. Similarly, the forces due to the upper and lower end stops are given by [16]

$$F_{\mathrm{u}} = \begin{cases} -k_{\mathrm{u}}(x-l_{\mathrm{u}}) & \text{if } x > l_{\mathrm{u}} \\ 0 & \text{else} \end{cases}, \tag{8}$$

$$F_{\mathrm{l}} = \begin{cases} -k_{\mathrm{l}}(x+l_{\mathrm{l}}) & \text{if } x < -l_{\mathrm{l}} \\ 0 & \text{else} \end{cases}, \tag{9}$$

where k_{u} and k_{l} are the equivalent stiffness values of the springs in the upper and lower end stops respectively. l_{u} and l_{l} are the distance of the two end stops from the vertical midpoint of the translator at equilibrium, as shown in Fig. 1. In [16], it is possible to find the values of l_{l}, l_{u}, k_{l}, k_{u}, k_{w}, k_{s}, m_{b}, m_{p}, F_0 and S_{w}.

Using (3)–(9), (1)–(2) have been expressed in the following non-linear state-space form

$$\dot{\boldsymbol{z}}(t) = \boldsymbol{A}\boldsymbol{z}(t) + \boldsymbol{B}\boldsymbol{u}(t,x) + \boldsymbol{B}\boldsymbol{w}(t) + \boldsymbol{B}\boldsymbol{l}(t,x,y), \tag{10}$$

The state, input, noise and non-linear vectors are given by

$$\boldsymbol{z} = \begin{bmatrix} y & \dot{y} & x & \dot{x} & \boldsymbol{x}_{\mathrm{ss}}^T \end{bmatrix}^T, \tag{11}$$

$$\boldsymbol{u} = \begin{bmatrix} 0 & -F_{\mathrm{em}}(x) \end{bmatrix}^T, \tag{12}$$

$$\boldsymbol{w} = \begin{bmatrix} F_{\mathrm{e}}(t) & 0 \end{bmatrix}^T, \tag{13}$$

$$\boldsymbol{l} = \begin{bmatrix} -F_{\mathrm{w}}(x,y) & F_{\mathrm{w}}(x,y) - F_0 + F_{\mathrm{u}}(x) + F_{\mathrm{l}}(x) \end{bmatrix}^T. \tag{14}$$

The state and input matrices are

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \boldsymbol{0}^T \\ -\frac{\rho g S_{\mathrm{w}}}{m_{\mathrm{b}}+m_\infty} & 0 & 0 & 0 & -\frac{\boldsymbol{C}_{\mathrm{ss}}}{m_{\mathrm{b}}+m_\infty} \\ 0 & 0 & 0 & 1 & \boldsymbol{0}^T \\ 0 & 0 & -\frac{k_{\mathrm{s}}}{m_{\mathrm{p}}} & 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{B}_{\mathrm{ss}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_{\mathrm{ss}} \end{bmatrix}, \tag{15}$$

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 \\ \frac{1}{m_{\mathrm{b}}+m_\infty} & 0 \\ 0 & 0 \\ 0 & \frac{1}{m_{\mathrm{p}}} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \tag{16}$$

C. Resistive Control of the Translator

The motions of the float and the translator, and thus ultimately the power absorption of the device, can be controlled through the electrical behaviour of the generator. In particular, the electromotive force is proportional to stator current I_{s} and active area A_{fac}

$$F_{\mathrm{em}} = k_\tau A_{\mathrm{fac}}(x) I_{\mathrm{s}}, \tag{17}$$

where k_τ is the generator torque constant. If the current is controlled (by power electronics) so that it is proportional to speed such as $I_{\mathrm{s}} = b\dot{x}$, with b being a constant, then (17) becomes

$$F_{\mathrm{em}} = k_\tau b A_{\mathrm{fac}}(x)\dot{x}, \text{ or } F_{\mathrm{em}} = \gamma A_{\mathrm{fac}}(x)\dot{x}, \tag{18}$$

where $\gamma = k_\tau b$ is the PTO damping coefficient. The active area, i.e. the overlap between stator and translator, is given by

$$A_{\mathrm{fac}} = \begin{cases} 0 & \text{if } |x| \geq 0.5(l_{\mathrm{p}} + l_{\mathrm{s}}) \\ 1 & \text{if } |x| \leq 0.5(l_{\mathrm{p}} - l_{\mathrm{s}}) \\ \left[0.5(l_{\mathrm{p}} + l_{\mathrm{s}}) - |x|\right]/l_{\mathrm{s}} & \text{else} \end{cases},$$

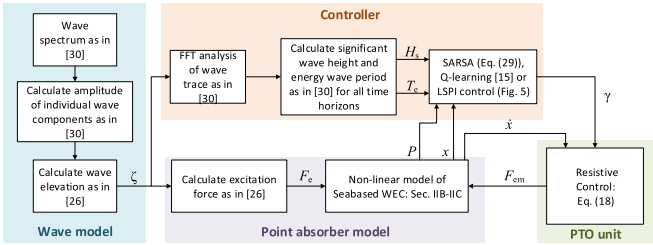

Fig. 2. Workflow diagram of the computer program used to simulate the Seabased WEC.

with l_p and l_s being given in [16]. The generated power is computed as

$$P(t) = F_\mathrm{em}(t)\dot{x}(t). \tag{19}$$

RL is employed to find the optimal PTO damping coefficient, γ, in each sea state for the maximization of the energy generation. The values of γ are assumed to be limited to 0–100 kNs/m. The upper limit corresponds approximately to the case of no load resistance in the experimental setting in [16].

D. Simulation System

Equation (10) has been discretized with a fourth-order Runge-Kutta scheme [28], and solved with a time step of 0.002 s. The controller is implemented as in (18). The workflow of the program is similar to that described in [15], and it is summarised in Fig. 2 with a block diagram.

III. Reinforcement Learning

In the RL framework [29], an agent takes an action a in state s, landing in a new state s' while observing a reward r. A Markov decision process is used to model the action selection depending on the value function $Q(s,a)$, which represents an estimate of the future reward. With time, the agent learns an optimal policy, π, that maximizes the total reward.

A. Least-Squares Policy Iteration

LSPI is a powerful, off-line, off-policy RL algorithm. Whilst still being model-free, the method, developed by [17], presents an efficient use of the samples (s,a,s',r), which results in a smaller learning time as compared with other strategies, such as Q-learning and SARSA. Additionally, it automatically incorporates function approximation for the action-value function. This means that the scheme is able to generalize for unseen states, thus further shrinking the convergence time. In particular, a linear architecture is used for the approximation of Q due to its simple implementation and ease of debugging and feature engineering [17]. In matrix notation, this is expressed as [18]

$$Q(s,a) \approx \boldsymbol{\phi}(s)^T \boldsymbol{\Theta}_{:,a}, \tag{20}$$

where $\boldsymbol{\Theta}$ is the weight matrix and $\boldsymbol{\phi}$ is the vector of arbitrary, linearly independent, usually non-linear basis functions, or features. $\boldsymbol{\Theta}_{:,a}$ indicates the a^th column of $\boldsymbol{\Theta}$, with $\boldsymbol{\Theta}$ having $|\mathcal{A}|$ columns, where $\mathcal{A}$ is the action space. $\boldsymbol{\Theta}$ and $\boldsymbol{\phi}$ have $J \ll |\mathcal{S}|$ rows, with $\mathcal{S}$ indicating the state space. Here, two basis functions

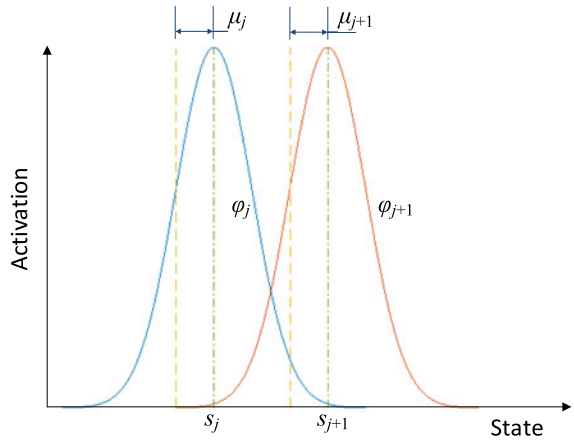

Fig. 3. Activation function of the RBFs as per (21).

input: $\boldsymbol{W}$: set of samples (s,a,r,s')
γ_d: discount factor
$\delta = 10^{-3}$: stopping criterion
π_0: initial policy, given as $\boldsymbol{\Theta}_0 = \mathbf{0}$
π: policy, or exploration strategy

- $\boldsymbol{\Theta}' \leftarrow \boldsymbol{\Theta}_0$
- **while** $||\boldsymbol{\Theta} - \boldsymbol{\Theta}'|| \geq \delta$:
  - $\Theta \leftarrow \boldsymbol{\Theta}'$
  - $\tilde{\boldsymbol{A}} \leftarrow \mathbf{0}$ ($J \times J$ matrix)
  - $\tilde{\boldsymbol{b}} \leftarrow \mathbf{0}$ (J vector)
  - **for** each $(s,a) \in \boldsymbol{W}$:
    - $\tilde{\boldsymbol{A}} \leftarrow \tilde{\boldsymbol{A}} + \boldsymbol{\phi}(s)\left(\boldsymbol{\phi}(s) - \gamma_d \boldsymbol{\phi}(s', \pi(s'))\right)^T$
    - $\tilde{\boldsymbol{b}} \leftarrow \tilde{\boldsymbol{b}} + \boldsymbol{\phi}(s) r$
    - $\boldsymbol{\Theta}'_{:,a} \leftarrow \tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{b}}$
- **return** $\boldsymbol{\Theta}$

Fig. 4. LSPI algorithm, adapted from [17].

types are used: tabular and radial. The tabular representation is the simplest and consists in assigning a separate weight for each state-action pair [18]. Hence, for discrete states, this corresponds to the exact representation $Q(s,a)$, although its size is equal to the state-action space ($J = |\mathcal{S}|$). Conversely, in RBFs, the feature activation decays continuously away from the state-action pair where the RBF is centred, s_j for RBF j, spanning many discrete states [18]

$$\phi_j(s) = \exp\left(-\frac{||s - s_j||^2}{2\mu_j}\right), \tag{21}$$

where μ_j indicates the bandwidth of RBF j. RBFs are shown graphically in Fig. 3.

LSPI consists of two main stages: policy evaluation (the critic) and policy improvement (the actor) [17]. LSPI is defined as off-line because the algorithm is trained using samples that have been previously recorded from observations of the environment. The algorithm is summarized in Fig. 4. The discount factor is set here to $\gamma_\mathrm{d} = 0.95$. The values of the weight matrix in (20) can be computed from

$$\tilde{\boldsymbol{A}}\boldsymbol{\Theta}_{:,a} = \tilde{\boldsymbol{b}} \tag{22}$$

for each action, where the tilde indicates a learned variable. The reader is referred to [17] for a full derivation of the equations for

$\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{b}}$, which are obtained from the least-squares fixed-point approximation.

B. Application of LSPI to the Control of WECs

Employing a time-averaged approach, at the start of each time-averaging period, or time horizon with duration H_{RL}, an action, which consists in a step change in the PTO damping coefficient, is selected following the current policy. The state is a combination of γ and the sea state, as given by the significant wave height H_{s} and energy wave period T_{e} [30]. Holding γ constant during H_{RL}, the reward is obtained as a function of the mean generated power, P_{avg}. The selection of a new action results in a new state, and the sample (s, a, s', r) is added to the sample set $\boldsymbol{W}$. After the collection of N_{s} samples, the policy is updated using the LSPI algorithm in Fig. 4. In the following sections it is possible to find an accurate description of the state and action spaces, the reward function and the exploration strategy.

1) State Space: Similarly to [15], the discrete state-space can be expressed as

$$\mathcal{S} = \left\{ s | s_{i,l,m} = (H_{\mathrm{s},i}, T_{\mathrm{e},l}, \gamma_m), \begin{array}{c} i = 1:I, \\ l = 1:L, \\ m = 1:M \end{array} \right\}. \quad (23)$$

As described in Section III-A, LSPI incorporates linear function approximation. With the tabular approach, $J = ILM$, i.e. there is an entry in $\boldsymbol{\Theta}$ for each state-action value, or the Q-table is exact. With RBFs, a smaller number of values can be used. In fact, for the control of WECs, a hybrid approach is used, where discrete sea states are still employed, while RBFs approximate the control variable. I and L are determined from the wave data at the deployment site, with steps of 1 m and 1 s being common for H_{s} and T_{e}, respectively [30].

For the tabular approach, $M = 11$ has been selected, with γ ranging from 0 to 100 kNs/m in steps of 10 kNs/m. For function approximation, 5 cases have been considered in order to study the influence of the number of kernels, or centres, and bandwidth on the learning behaviour of LSPI with RBFs. In Table I, it is possible to see the distance between kernels $\delta_c = s_j - s_{j-1}$ and bandwith μ for each case as given in (21). The first kernel is always sited at $\gamma = 0$ kNs/m.

2) Action Space: The action space is defined as

$$\mathcal{A} = \{a | (-\Delta\gamma, 0, +\Delta\gamma)\}, \quad (24)$$

where $\Delta\gamma = \gamma_{m+1} - \gamma_m$. However, the states corresponding to the maximum and minimum of the PTO damping coefficient are constrained to two actions in order to avoid going beyond the RL state space limits.

3) Reward Function: The same reward function as in [15] is employed and the reader is referred there for a more detailed explanation. A penalty $p = -1$ is returned if the constraints $\max(x) > l_{\mathrm{u}} + l_{\mathrm{e,u}}$ or $\min(x) < -(l_{\mathrm{l}} + l_{\mathrm{l,u}})$ are exceeded during the time horizon h. This enables the algorithm to learn to avoid actions that will result in possible damage to the device. Hence, the reward function is expressed as

$$r = \begin{cases} \left[\frac{\boldsymbol{m}(s_h)}{\max_{s''=o:p} \boldsymbol{m}(s'')} \right]^u & \text{if constraints met} \\ p & \text{otherwise} \end{cases}. \quad (25)$$

The entries of the vector $\boldsymbol{m}$, whose size is equal to the total number of discrete states $|\mathcal{S}|$ (valid for both tabular and RBF methods), correspond to the average of up to 10 values of $P_{\mathrm{avg}}/H_{\mathrm{s}}^2$ that are stored for each discrete state, with older values being overwritten by new ones once 10 values are registered. The indices o and p ensure that the maximization in (25) is performed only over the values of $\boldsymbol{m}$ corresponding to the current sea state, as given by H_{s} and T_{e}.

The power u must be an odd number to prevent rectifying negative power values (in the case of reactive control). The higher the value, the closer the cost function is to returning 1 for the optimal control variable and 0 for all other settings in each sea state. Here, $u = 25$ has been used in order to aid convergence in irregular waves.

4) Exploration Strategy: An ϵ-greedy policy [29] selects the action at the start of each time horizon h

$$a = \begin{cases} \arg\max_{a' \in \mathcal{A}} Q(s_h, a') & \text{with probability } 1 - \epsilon_h \\ \text{random action} & \text{with probability } \epsilon_h \end{cases}, \quad (26)$$

where ϵ_h is the exploration rate. $\max_{a' \in \mathcal{A}} Q(s_h, a')$ represents the maximum action-value (i.e. a measure of the expected reward) for the current state over all actions, with the action-value function being given by the mapping in (20). This term represents the selection of the action that results in maximum expected total reward starting from the current state.

Greater exploration is desired at the start of RL control, while the greedy action, i.e. such that it maximises the value function, is preferred as the learned policy improves. Thus, the exploration rate is obtained as

$$\epsilon_h = \begin{cases} \epsilon_0 & \text{if } \boldsymbol{N}(s_h) \leq N_\epsilon \\ \epsilon_0 / \sqrt{\boldsymbol{N}(s_h) - N_\epsilon} & \text{if } \boldsymbol{N}(s_h) > N_\epsilon \end{cases}, \quad (27)$$

with $\boldsymbol{N}(s_h)$ indicating the number of visits to the current discrete state (hence, valid for both tabular and RBF approaches). $N_\epsilon = 5$ is the minimum number of encounters for random exploration, and the initial exploration rate is set to $\epsilon_0 = 0.5$.

C. Algorithm

The proposed LSPI algorithm for the resistive control of the stator can be seen in Fig. 5. After initializing all variables, the algorithm is run continuously until the device is disconnected, e.g. due to maintenance. During each time horizon h, the policy is applied in order to select a suitable action based on the encountered sea state, mean generated power and maximum translator displacement. Furthermore, at the end of each horizon, the current state, action, next state and reward are sampled and added to $\boldsymbol{W}$. Due to the finite memory of the controller computer, a specified number of samples can be stored, say 10^6. Therefore, new samples will be stored only if they have not been recorded before, with a difference greater than 10^{-3} being acceptable for the reward. Once the memory limit is reached, older values will need to be overwritten, ensuring the sample range is broad, i.e.

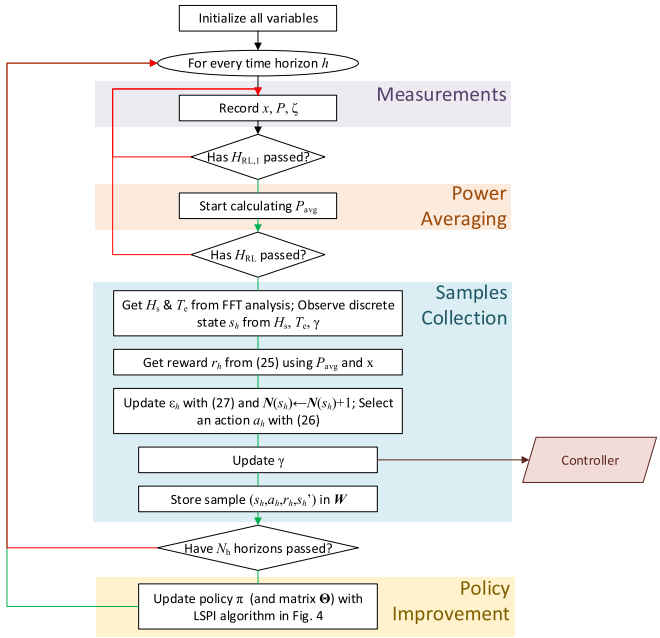

Fig. 5. Flowchart of the LSPI algorithm for the resistive control of the stator.

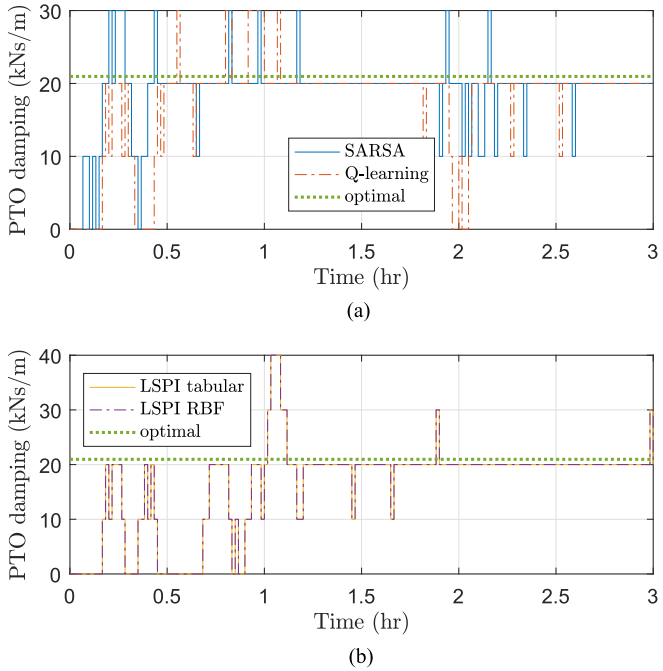

Fig. 6. PTO damping coefficient selected by different RL control strategies as compared with the optimal value in regular waves with unit amplitude and $T = 6$ s starting from $\gamma = 0$ kNs/m.

accounting for the different sea states and values of the PTO damping coefficient.

As shown in Fig. 5, the policy is improved using the LSPI algorithm in Fig. 4 every $N_h = 40$ time horizons. This operation can be performed off-line on separate computing cores so as to reduce the computational effort and ensure the on-line implementation is feasible.

A time horizon duration $H_{\mathrm{RL}} = 10T$ has been chosen in regular waves, with T being the wave period, while $H_{\mathrm{RL}} = 150$ s in the analysed irregular waves because a JONSWAP spectrum is used. This selection is based on a compromise between a fast response and a stable algorithm. Irregular waves in particular require a longer duration of the power averaging process due to their stochastic nature. If a wider-banded wave spectrum is adopted, the horizon length should be increased. Additionally, this process is started only after $H_{\mathrm{RL},1} = 0.4H_{\mathrm{RL}}$ so as to remove the influence of the transient effects associated with the change in load resistance.

D. Q-Learning and SARSA

The performance of LSPI in the control of the WEC is compared with Q-learning and SARSA. SARSA, which stands for *state-action-reward-state-action*, and Q-learning are on-line schemes that rely on discrete states and actions [29]. Hence, at each step, they update the Q-table with the following equations [29]

$$\boldsymbol{Q}(s,a) \leftarrow \boldsymbol{Q}(s,a_)+\alpha\left[r+\gamma_{\mathrm{d}}\max_{b\in\mathcal{A}}\boldsymbol{Q}(s',b)-\boldsymbol{Q}_(s,a)\right], \tag{28}$$

$$\boldsymbol{Q}(s,a) \leftarrow \boldsymbol{Q}(s,a_)+\alpha\left[r+\gamma_{\mathrm{d}}\boldsymbol{Q}(s',a')-\boldsymbol{Q}_(s,a)\right], \tag{29}$$

for Q-learning and SARSA respectively. α is the learning rate and a' the action applied in the future state. Hence, it is clear that the main difference between the two algorithms is that while SARSA is on-policy, i.e. it updates the value function based on the policy it will follow, Q-learning is off-policy, i.e. the update is based on the maximum possible Q-value in the new state [29]. The application of Q-learning to the resistive control of WECs is described in a previous publication [15], with SARSA presenting an almost identical implementation. Hence, the reader is referred to [15] for details. The initial learning rate is set here to 0.4.

IV. Simulation Results

A. Regular Waves

The behaviour of SARSA, Q-learning and LSPI has been assessed against the optimal PTO damping coefficient, which has been calculated using the Matlab optimization function *fmincon* in each sea state. This is to provide a benchmark of the control variable that results in the maximum mean generated power.

Regular waves of unit amplitude and a wave period of 6 s have been analysed first, with a wave trace lasting 3 hours. Two different starting points have been selected, namely $\gamma = 0$ and $\gamma = 100$ kNs/m, as shown in Figs. 6 and 7, respectively. For the RBFs, $\delta_c = 10$ kNs/m and $\mu = 10$ kNs/m, i.e. an almost tabular approach has been used. For each figure, the same seed number has been set to the random number generator for all algorithms, selecting a particularly unfavourable number for Fig. 7 in order to assess the convergence properties under difficult conditions.

In Fig. 8, it is possible to see the behaviour of the LSPI algorithm for the RBF settings in Table I, when the starting

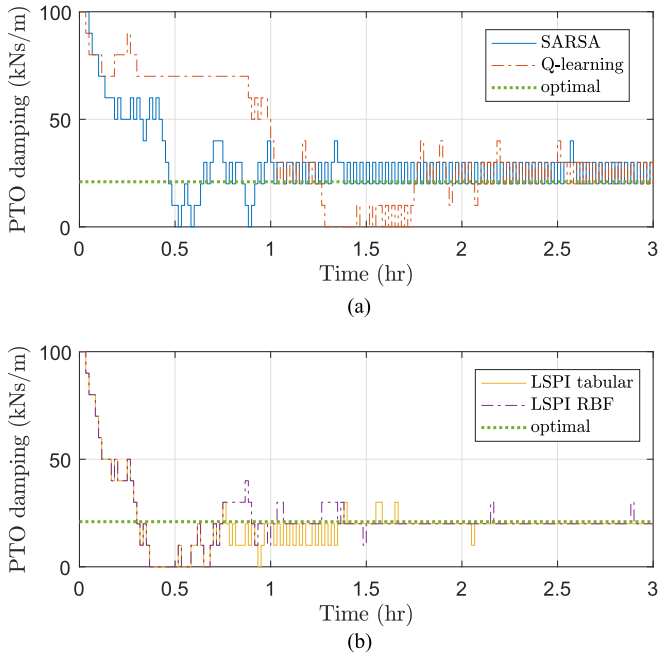

Fig. 7. PTO damping coefficient selected by different RL control strategies as compared with the optimal value in regular waves with unit amplitude and $T = 6$ s starting from $\gamma = 100$ kNs/m.

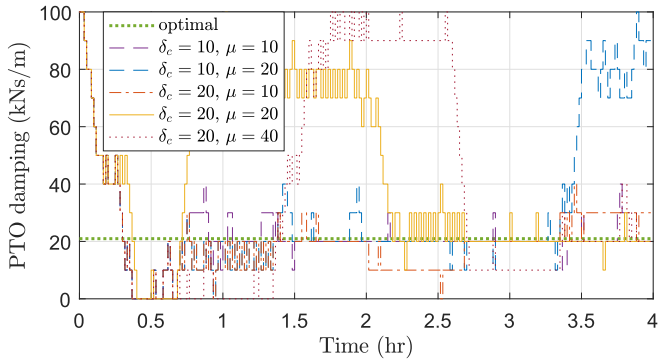

Fig. 8. PTO damping coefficient selected by the LSPI algorithm with different RBF settings in regular waves with unit amplitude and $T = 6$ s. The values of δ_c and μ are in kNs/m.

value of the PTO coefficient is $\gamma = 100$ kNs/m. For all runs, the same seed values is used as in Fig. 7. A longer wave trace lasting 4 hours is employed.

The mean generated power corresponding to the run with LSPI with RBFs and $\delta_c = 10$ kNs/m and $\mu = 10$ kNs/m in Figs. 7(b) and 8 is plotted in Fig. 9.

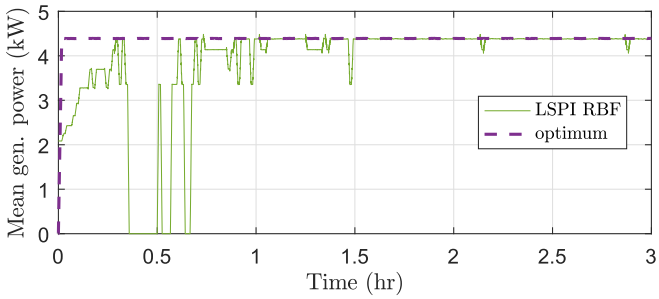

Fig. 9. Mean generated power for the run with LSPI with RBFs and $\delta_c = 10$ kNs/m and $\mu = 10$ kNs/m in Figs. 7(b) and 8.

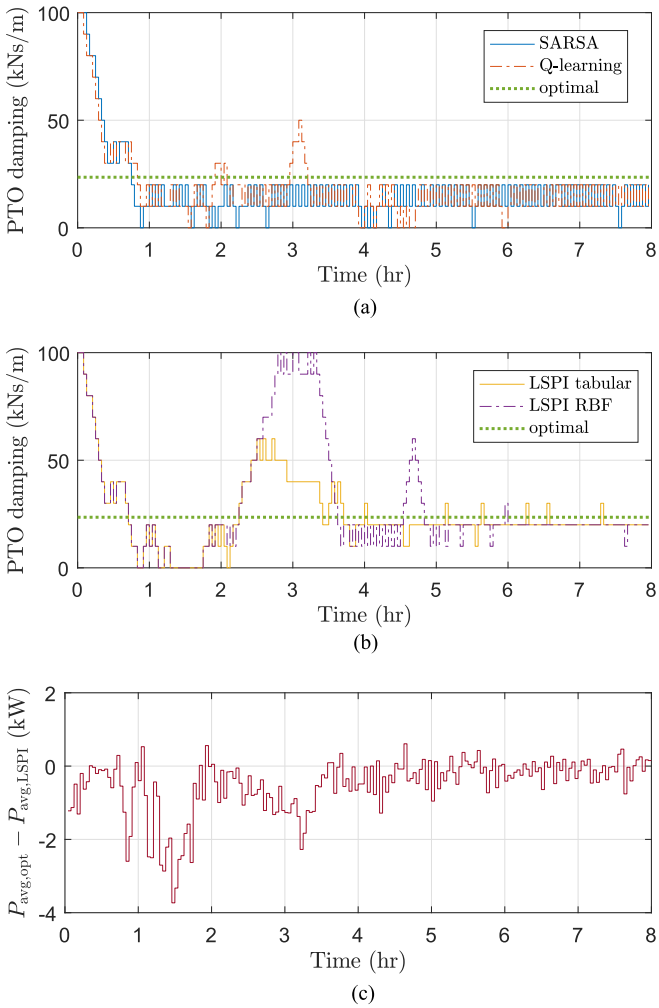

Fig. 10. PTO damping coefficient selected by different RL control strategies as compared with the optimal value in irregular waves with $H_s = 2$ m and $T_e = 6$ s and a JONSWAP spectrum starting from $\gamma = 100$ kNs/m (a)–(b). (c) Shows the difference in the mean generated power for the optimum ($P_{\mathrm{avg,opt}}$) and the case of LSPI with RBFs ($P_{\mathrm{avg,LSPI}}$).

B. Irregular Waves

In irregular waves, an 8-hour long wave trace with $H_s = 2$ m and $T_e = 6$ s with a JONSWAP spectrum [30] has been analysed, typical of the Lysekil testing site [31]. In Fig. 10(a) and (b), the learning behaviours of the three control algorithms are shown, with the same setting being used for LSPI with RBFs as in Fig. 7 throughout this section. The difference in mean generated power between LSPI with RBFs and the optimal control setting is shown in Fig. 10(c).

Nevertheless, real sea states actually last between 0.5 to 6 hours [30]. Therefore, in order to prove that RL is able to deal with changing sea states, the control is tested in an additional 12-hour-long wave trace composed of the alternation of two sea

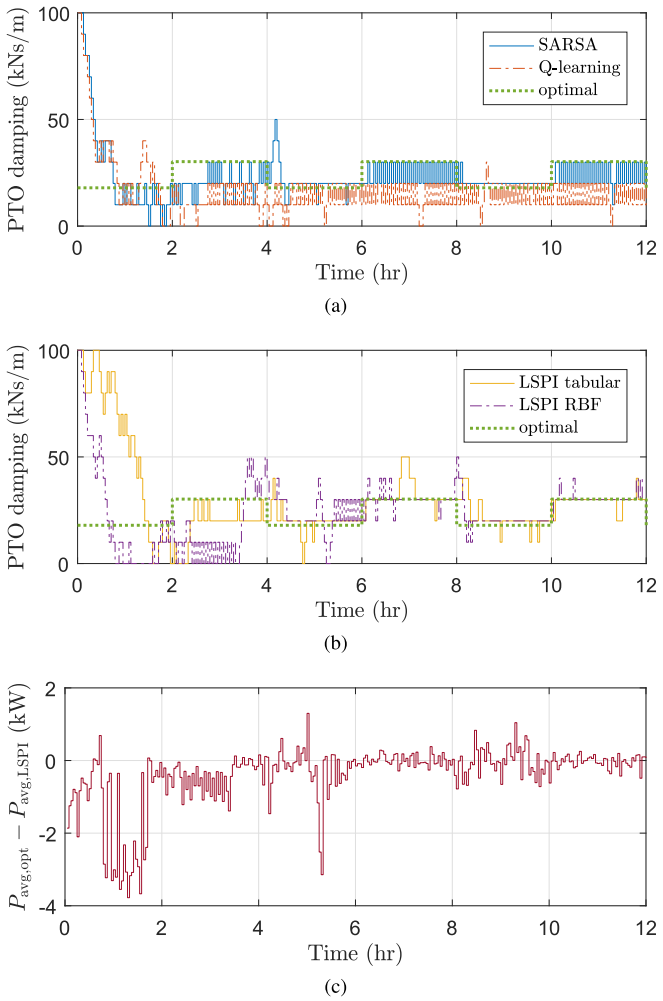

Fig. 11. PTO damping coefficient selected by different RL control strategies as compared with the optimal value in irregular waves with two alternating sea states (JONSWAP spectra with $H_\mathrm{s} = 2$ m and $T_\mathrm{e} = 5$ s, and $H_\mathrm{s} = 1$ m and $T_\mathrm{e} = 6$ s) starting from $\gamma = 100$ kNs/m (a)–(b). (c) Shows the difference in the mean generated power for the optimum ($P_{\mathrm{avg,opt}}$) and the case of LSPI with RBFs ($P_{\mathrm{avg,LSPI}}$).

states, so that $I = L = 2$. Both have a JONSWAP spectrum and last for two hours before changing. The first one corresponds to $H_\mathrm{s} = 2$ m and $T_\mathrm{e} = 5$ s, while the second one has $H_\mathrm{s} = 1$ m and $T_\mathrm{e} = 6$ s. Fig. 11(a) and (b) shows the learning behaviour of the three RL algorithms. In Fig. 11(c), the difference in mean power between LSPI with RBFs and the optimal control setting in each sea state can be seen.

Furthermore, although RL is expected to result in adaptive control, as it is model-independent [32], this was not proven in the previous work on the control of WECs [15]. Hence, a simple example is treated here to show the adaptivity of RL to possible marine growth effects. Bio-fouling is expected to affect the dynamics of the system mainly through an increase in its inertia and especially drag force. However, in this simple model, the viscous drag force is not considered. Hence, we treat the case of a sudden increase in the radius and draught of the floater to 1.75 m and 0.5 m, respectively (from 1.5 m and 0.4 m, respectively, in [16]). These values have been assumed, as they result in a significant change in the optimal damping coefficient in the analysed sea state. A full sensitivity analysis of the power absorption and control of the device to the variations in floater design as well as a realistic treatment of marine growth effects go beyond the scope this study. The changes in the floater design result in an increase in its mass (m_b) of 2032.7 kg, in its surface area of 2.5525 m^2 (note that these values are relative to the values in [16]) as well as changes in the radiation approximation state-space system, with the radiation coefficients being computed with WAMIT.

The same sea state as in Fig. 10 is used in this simple example, whereas the new geometry of the floater is employed. In particular, a simulation is initialized with the final values of Fig. 10(a) and (b) being set for each RL strategy. Additionally, the same values of the $\boldsymbol{m}$ vector have been kept for each scheme. This corresponds to initializing the reward function to incorrect values for each discrete damping coefficient. For this reason, the exploration rate (as well as the learning rate for Q-learning and SARSA) is reinitialized with the same settings as in Section III-B4.

In Fig. 12(a) and (b), the learning behaviours of the three control algorithms are shown. The difference in mean generated power between LSPI with RBFs and the optimal control setting is shown in Fig. 12(c).

The computational time of the algorithm run at the start of each time horizon has been less than 0.06 s on an i7 processor with 16Gb RAM in all simulations run here. As this time is proportional to the number of states, if, say, 100 sea states were to be used, the computational time would increase to 0.3 s. Hence, a practical implementation is realistic, particularly considering the much longer time horizon duration.

V. Discussion

A. Regular Waves

In this work, we define RL algorithms to have converged towards a policy once the same PTO damping coefficient is selected for longer than an hour. However, within the short duration of the analysed wave traces, the exploration rate does not fully decay. Hence, the definition of convergence is extended to include a maximum of up to 5 distinct deviations from the mean value of the selected γ within the one-hour period, which may be due to random actions being adopted.

In Fig. 6, it can be seen that all algorithms learn the optimal PTO damping coefficient within 2.5 hours, with subsequent wiggles, especially visible for Q-learning and SARSA, mainly due to the exploration rate not having fully decayed. This fast learning is because this is a benign case, with the optimal value of γ being very close to the starting PTO damping coefficient, thus requiring little exploration before finding the optimum. Conversely, Fig. 7 represents a more challenging scenario for the RL algorithms. In particular, SARSA and Q-learning are unable to converge to the optimal policy, and learn a suboptimal policy instead, which results in less energy absorption than the optimal policy. This problem could be solved with a slower decay in the exploration and learning rates, which would cause learning to be smoother, but also slower. This behaviour is

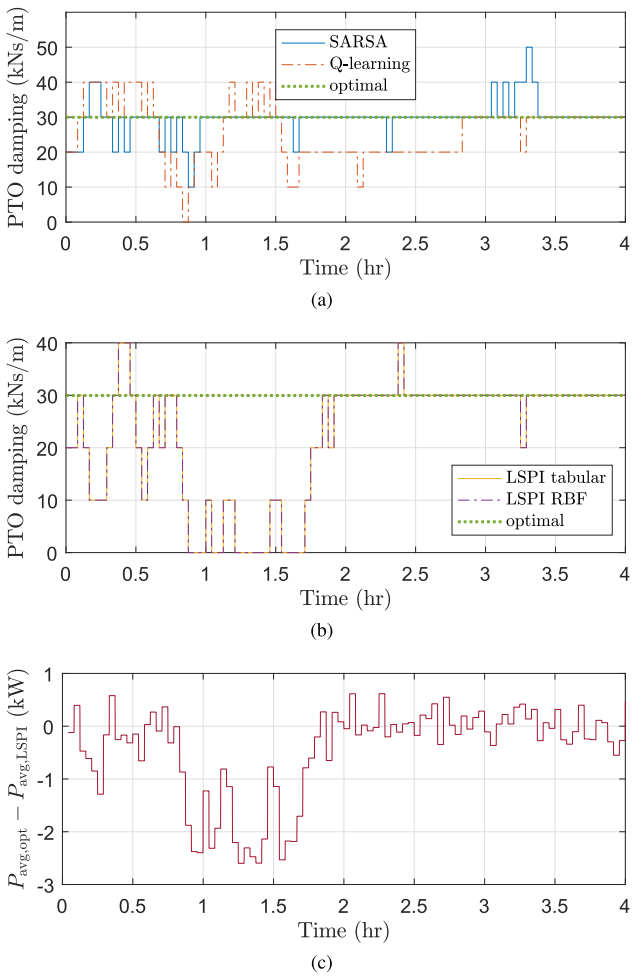

Fig. 12. PTO damping coefficient selected by different RL control strategies as compared with the optimal value for the new floater geometry in irregular waves with $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 6$ s and a JONSWAP spectrum. The initial conditions are set based on the final settings of Fig. 10(a) and (b), respectively. (c) Shows the difference in the mean generated power for the optimum ($P_{\mathrm{avg,opt}}$) and the case of LSPI with RBFs ($P_{\mathrm{avg,LSPI}}$).

particularly worrying in the case of extreme waves because if this oscillation occurs on the boundary of the feasible damping coefficient envelope to prevent excessive displacements, it could lead to failure. Conversely, LSPI with both tabular and radial basis functions learns the optimal policy within 2.5 hours in regular waves in Fig. 7(b).

Comparing the behaviour of LSPI with tabular features and RBFs with $\delta_c = 10$ kNs/m and $\mu = 10$ kNs/m in Figs. 6(b), 7(b) and 10(b), the two approaches almost completely match, with RBFs actually resulting in a stabler behaviour in regular waves and greater exploration in irregular waves. This is expected because almost the same number of kernels as discrete states are used, with the bandwidth spanning the space between discrete states. In Fig. 4, decreasing the number of kernels was expected to result in faster learning because the RBFs are expected to generalise the shape of the Q-function for unseen states and actions [18]. In fact, this is not the case, with LSPI with RBFs with $\delta_c = 20$ kNs/m (thus half as many kernels) and $\mu = 20$ kNs/m taking longer to learn the optimal policy. Increasing the bandwidth of RBFs also augments the confusion in the controller, as the overlap between distinct RBFs is increased spanning multiple γ values, thus causing the algorithm to diverge from the optimal policy. These counter-intuitive results are believed to be due to the small number of discrete states used, with many more features being typical for standard RL problems [17]. Hence, the use of 5 or less RBFs incurs in an underfitting problem, i.e. using too coarse a model to fit the Q-function. A minimum of 10 RBFs is recommended for the control of WECs with LSPI. Additionally, setting the bandwidth to match the distance between kernels seems to provide best behaviour. Nevertheless, designing RBFs features needs care, and it is likely to be device-specific.

B. *Irregular Waves*

Q-learning and SARSA are similarly unable to converge towards the optimal policy in irregular waves as well, as shown in Figs. 10 and 11. Again, this is an indicator that the exploration and learning rates should be decreased more slowly for these algorithms, thus resulting in longer learning times. Conversely, LSPI with both tabular and radial basis functions is able to learn the optimal policy in less than 6 hours in each sea state, despite some wiggles owing to the exploration rate not having decayed fully yet in Fig. 11(b). In particular, the learning time is lower than the 12 hours required by Q-learning for convergence in irregular waves in [15], where a more benign linear WEC model was used for validation. This diminished convergence time is mainly due to the shorter time-averaging horizon length employed in this study and, especially, the superior capacity of LSPI to learn using a small number of observations [17]. Furthermore, as shown in Fig. 11(b), LSPI is able to pick up learning in a specific sea state from where it left off the last time the controller was in that sea state. This is a fundamental consideration for a realistic application, since actual sea states usually last for a shorter time than 6 hours [30].

As the Seabased device is tested in the Skagerrak strait [16], a JONSWAP spectrum is appropriate due to its bounded, shallow-water nature [30]. However, a JONSWAP spectrum is a single-peaked spectrum with a relatively narrow frequency range [30]. This means that energy is contained mainly in a region close to the peak wave period. As a result, determining the optimal PTO damping coefficient for each sea state is simpler than for wider-banded wave spectra, such as Bretschneider or even double-peaked spectra. Although RL is expected to find the global optimum [29], the learning process would be expected to take longer if the latter spectra were used: a longer time horizon length would be necessary. In particular, a double-peaked spectrum would cause significant challenges to the convergence behaviour. This will be the focus of future studies.

Being model-free, RL is proven to be able to adapt to changes in the dynamics of the WEC in Fig. 12. Even though the reward function is initialized with the wrong values, RL is able to converge towards the optimal PTO damping coefficient with all three analysed algorithms. However, it is important to note that this is possible because the exploration rate is reset after the change of the system dynamics. Therefore, during operation of

a WEC, it is necessary to reset the exploration rate after specific time intervals, say yearly, in order to pick up any possible changes in the device response.

VI. Conclusion

An efficient RL algorithm has been suggested for the control of a WEC, with its performance being compared with Q-learning and SARSA. In particular, a non-linear model of the dynamics of the Seabased point absorber, validated in a previous study, has been used as a test case. As expected, despite the system non-linearities, all control schemes are able to find the optimal PTO damping coefficient from a random start in regular waves because of their model-free nature. However, if the algorithms are started with particularly unfavourable conditions, only LSPI is able to converge within 2.5 hours, with higher learning and exploration rates being required for Q-learning and SARSA to converge. Unexpected results have been found in the study of RBFs as features for function approximation with LSPI: a smaller number of RBFs than discrete states does not correspond to faster learning time. This is because a very small number of discrete states has been employed, with the few RBF kernels resulting in underfitting. Hence, although RBFs should be preferred over tabular features as they presented a stabler behaviour, their number should be high enough to prevent underfitting, thus meaning that their design is likely to be specific to the device dynamics.

In irregular waves, LSPI learns the optimal policy within 6 hours starting from unfavourable conditions, thus proving its superior capacity of learning from a limited set of observations. The same behaviour is observed when the controller is tested in two sea states, alternating every 2 hours. Finally, RL is shown to converge towards a new optimal policy after changing the floater geometry, with the controller still being initialized with the reward function valid for the older system. This proves the adaptive nature of RL control, supporting its ability to account for changes in the system dynamics, e.g. due to marine biofouling.

References

[1] K. Gunn and C. Stock-Williams, "Quantifying the potential global market for wave power," in *Proc. 4th Int. Conf. Ocean Eng.*, 2012, pp. 1–7.

[2] A. F. D. O. Falcão, "Wave energy utilization: A review of the technologies," *Renewable Sustain. Energy Rev.*, vol. 14, no. 3, pp. 899–918, 2010.

[3] S. H. Salter, J. R. M. Taylor, and N. J. Caldwell, "Power conversion mechanisms for wave energy," *Proc. Inst. Mech. Eng. M.*, vol. 216, no. 1, pp. 1–27, 2002.

[4] J. V. Ringwood, G. Bacelli, and F. Fusco, "Energy-maximizing control of wave-energy converters: The development of control system technology to optimize their operation," *IEEE Control Syst. Mag.*, vol. 34, no. 5, pp. 30–55, Oct. 2014.

[5] A. J. Nambiar, D. I. M. Forehand, M. M. Kramer, R. H. Hansen, and D. M. Ingram, "Effects of hydrodynamic interactions and control within a point absorber array on electrical output," *Int. J. Marine Energy*, vol. 9, pp. 20–40, 2015.

[6] A. Babarit and A. H. Clément, "Optimal latching control of a wave energy device in regular and irregular waves," *Appl. Ocean Res.*, vol. 28, no. 2, pp. 77–91, 2006.

[7] A. Babarit, M. Guglielmi, and A. H. Clément, "Declutching control of a wave energy converter," *Ocean Eng.*, vol. 36, no. 12/13, pp. 1015–1024, 2009.

[8] F. Fusco and J. V. Ringwood, "A simple and effective real-time controller for wave energy converters," *IEEE Trans. Sustain. Energy*, vol. 4, no. 1, pp. 21–30, Jan. 2013.

[9] T. K. A. Brekken, "On model predictive control for a point absorber wave energy converter," in *Proc. IEEE Trondheim PowerTech*, 2011, pp. 1–8.

[10] G. Li and M. R. Belmont, "Model predictive control of sea wave energy converters—Part I: A convex approach for the case of a single device," *Renewable Energy*, vol. 69, pp. 453–463, 2014.

[11] M. Richter, O. Sawodny, M. E. Magaña, and T. K. A. Brekken, "Power optimisation of a point absorber wave energy converter by means of linear model predictive control," *IET Renewable Power Gener.*, vol. 8, no. 2, pp. 203–215, 2014.

[12] J. C. Henriques, L. M. Gato, A. F. Falcão, E. Robles, and F. X. Faÿ, "Latching control of a floating oscillating-water-column wave energy converter," *Renewable Energy*, vol. 90, pp. 229–241, 2016.

[13] F. Fusco and J. V. Ringwood, "Hierarchical robust control of oscillating wave energy converters with uncertain dynamics," *IEEE Trans. Sustain. Energy*, vol. 5, no. 3, pp. 958–966, Jul. 2014.

[14] M. P. Schoen, J. Hals, and T. Moan, "Wave prediction and robust control of heaving wave energy devices for irregular waves," *IEEE Trans. Energy Convers.*, vol. 26, no. 2, pp. 627–638, Jun. 2011.

[15] E. Anderlini, D. I. M. Forehand, P. Stansell, Q. Xiao, and M. Abusara, "Control of a point absorber using reinforcement learning," *Trans. Sustain. Energy*, vol. 7, no. 4, pp. 1681–1690, 2016.

[16] M. Eriksson, R. Waters, O. Svensson, J. Isberg, and M. Leijon, "Wave power absorption: Experiments in open sea and simulation," *J. Appl. Phys.*, vol. 102, no. 8, 2007, Art. no. 084910.

[17] M. G. Lagoudakis and R. Parr, "Least-squares policy iteration," *J. Mach. Learn. Res.*, vol. 4, pp. 1107–1149, 2003.

[18] A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, and J. P. How, "A tutorial on linear function approximators for dynamic programming and reinforcement learning," *Found. Trends Mach. Learn.*, vol. 6, no. 4, pp. 375–451, 2013.

[19] O. Danielsson, "Wave energy conversion: Linear synchronous permamnet magnet generator," Ph.D. dissertation, , Dept. Eng. Sci., Uppsala Univ., Uppsala, Sweden, 2006.

[20] M. Eriksson, "Modelling and experimental verification of direct drive wave energy conversion," Ph.D. dissertation, , Dept. Eng. Sci., Uppsala Univ., Uppsala, Sweden, 2007.

[21] R. Waters, "Energy from ocean waves," Ph.D. dissertation, Dept. Eng. Sci., Uppsala Univ., Uppsala, Sweden, 2008.

[22] M. Stalberg, R. Waters, O. Danielsson, and M. Leijon, "Influence of generator damping on peak power and variance of power for a direct drive wave energy converter," *J. Offshore Mech. Arctic Eng.*, vol. 130, no. 3, pp. 1–4, 2008.

[23] E. Lejerskog, C. Boström, L. Hai, R. Waters, and M. Leijon, "Experimental results on power absorption from a wave energy converter at the Lysekil wave energy research site," *Renewable Energy*, vol. 77, pp. 9–14, 2015.

[24] M. Eriksson, J. Isberg, and M. Leijon, "Theory and experiment on an elastically moored cylindrical buoy," *IEEE J. Ocean. Eng.*, vol. 31, no. 4, pp. 959–963, Oct. 2006.

[25] D. Forehand, A. E. Kiprakis, A. Nambiar, and R. Wallace, "A Bi-directional wave-to-wire model of an array of wave energy Converters," *IEEE Trans. Sustain. Energy*, vol. 7, no. 1, pp. 118–128, Jan. 2016.

[26] J. Falnes, *Ocean Waves and Oscillating Systems*, Paperback ed. Cambridge, U.K: Cambridge Univ. Press, 2005.

[27] G. F. Franklin, J. D. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, 6th ed. Upper Saddle River, NJ, USA: Pearson, 2008.

[28] I. H. Hutchinson, *A Student's Guide to Numerical Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning*, Hardcover ed. Cambridge, MA, USA: MIT Press, 1998.

[30] L. H. Holthuijsen, *Waves in Oceanic and Coastal Waters*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[31] R. Waters, J. Engström, J. Isberg, and M. Leijon, "Wave climate off the Swedish west coast," *Renewable Energy*, vol. 34, no. 6, pp. 1600–1606, 2009.

[32] F. Lewis, D. Vrabie, and K. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

Enrico Anderlini received the M.Eng. degree in naval architecture from the University of Southampton, Southampton, U.K., in 2013. He is currently working toward the D.Eng. degree in offshore renewable energy at IDCORE, a partnership of the Universities of Edinburgh, Exeter and Strathclyde, U.K.

His research interests include marine hydrodynamics, control of wave energy converters, and machine learning.

David I. M. Forehand received the B.Sc. (Hons.) degree in mathematics in 1990 and the Ph.D. degree in numerical modeling of free-surface waves from the University of Edinburgh, U.K., in 1999 and the M.Sc. degree in applied mathematics from the University of Oxford, Oxford, U.K., in 1993. He is a Research Associate with the Institute for Energy Systems, University of Edinburgh, U.K. His research interests include nonlinear engineering dynamics and the numerical hydrodynamic modeling of wave energy converters and floating bodies.

Elva Bannon received the B.Eng. (Hons.) degree in mechatronic engineering and the M.Eng. degree in advanced engineering from Dublin City University, Dublin, Ireland. She then worked in WEC Technology Research and Development focussing on tank testing, modeling, and simulation before taking up a post as a Senior Research Engineer for Wave Energy Scotland.

Mohammad Abusara received the B.Eng. degree from Birzeit University, Palestine, Birzeit, in 2000, and the Ph.D. degree from the University of Southampton, Southampton, U.K., in 2004, both in electrical engineering.

He is currently a Senior Lecturer in Renewable Energy at the University of Exeter, U.K. He has more than ten years of industrial experience with Bowman Power Group, Southampton, U.K., in the field of research and development of digital control of power electronics. During his years in industry, he led the development of a number of commercial products that include grid and parallel connected inverters, microgrid, dc/dc converters for hybrid vehicles, and sensorless drives for high-speed permanent magnet machines.