Control of a Wave Energy Converter using Reinforcement Learning

José Carlos Mota Trigueiro
jose.trigueiro@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

December 2021

Abstract

The development of control strategies that maximize power generation in Wave Energy Converters is fundamental in making the exploitation of sea waves an economically viable element of the energy mix. Classical, model-based control techniques have significant limitations in achieving this goal, due to their dependency on modelling accuracy and inability to adapt to changing system dynamics over time. In this thesis a control scheme based on Deep Reinforcement Learning (DRL) is presented, using a *MATLAB* and *Simulink* model of the Mutriku Oscillating Water Column plant as a training environment. This controller acts on the power take-off electromagnetic torque and relief valve aperture simultaneously, and exclusively uses data measured in the plant itself as observation signal, without requiring an external measuring tool for estimation of the sea state. Three different agent architectures are trained and tested: Deep Deterministic Policy Gradient (DDPG), Twin Delayed DDPG (TD3) and Soft Actor-Critic (SAC). Using as a baseline a power control law developed by previous authors, these agents are compared in terms of their expected yearly electric power production. The black box behaviour of the controller is also analysed, in an effort to gain insight into the type of learned control law it implemented.

Keywords: wave energy converter, oscillating water column, Mutriku, power take-off control, deep reinforcement learning

1. Introduction

The need for the mitigation of man-made climate change means that one of humanity's most difficult challenges for the 21st century is performing a successful transition from fossil fuel-based to renewable energy generation. An alternative to fulfill this demand is the widespread adoption of renewable energy sources, including the exploitation of ocean energy, particularly wave energy. There is currently only 2.31 MW of installed wave energy in the world, although it has an estimated theoretical potential for yearly energy production of 29500 TWh [14], enough to cover the world's electricity needs in 2019 of 26 730 TWh [13].

Multiple methods to harness energy from the waves were developed, the most common being oscillating body, Oscillating Water Column (OWC) and overtopping devices [14]. This work focuses on OWC devices, which extract energy through the compression of air in an air chamber caused by the wave oscillations , driving a power take off system constituted by a self-rectifying air turbine (Wells or biradial) and an electrical generator.

The main objective of this work is the development of a model-free control strategy for the Mutriku OWC plant, using data available at the plant, through the use of Deep Reinforcement Learning techniques. This control scheme should be adaptable to different sea states and respect the safety constraints of the plant while maximising power production, acting on two control variables: the generator torque and the aperture of a relief valve.

1.1. State of the Art in OWC control

The control problem in OWC devices may be formulated with multiple objectives in mind, such as maximizing electrical power, keeping the air turbine close to its optimal operating point or minimising undesirable events such as turbine stalling. Multiple strategies have been used aiming to achieve these objectives, including classical control using both frequency and time domain models and modern control, based on computational intelligence and the use of data.

Frequency domain control relies on approximating a set of optimality conditions for maximum power absorption in magnitude and phase [6]. To achieve these conditions, the impedance of the PTO system must match the mechanical impedance of the OWC for the frequencies found in the waves. This behaviour is commonly approximated through the latching of the WEC oscillation [11].

In the time domain, two main modes of control are identified for OWC devices: turbine rotational speed control, and airflow control.

Rotational speed control is performed by regulating the electromagnetic torque of the generator,

where the most common control law uses a cubic relation between rotation velocity and power in the generator, emulating the turbine performance curve [12].

Airflow control is performed through the use of valves both in series and in parallel. Valves in series have been shown to increase power generation while limiting rotation velocity in highly energetic sea states [5]. Safety valves in series with the air turbine have mostly been used to cut off air flow from the turbine when it reaches a threshold rotation velocity [12].

The main topic of this work, Reinforcement Learning (RL), has also been approached before as a strategy to control WEC devices, mostly of the oscillating body type, using tabular [2] or deep [3] RL, although some research has also been done on RL control of OWC's [7].

2. Reinforcement Learning Algorithms

Reinforcement Learning (RL) methods are based on an interaction between an agent, who acts as a decision maker and learner, and an environment, characterised by a set of observations available to the agent. When the agent performs an action, the environment's state changes, altering the possible future actions and respective outcomes, and providing numerical reward. The objective of the learning process is to maximise the expected future rewards of all subsequent action-state combinations. To achieve this goal, the agent must balance the exploitation of actions that have previously achieved a high reward, and the exploration of new, unknown actions.

The problem structure behind RL may be formulated in the framework of Markov Decision Processes (MDP's), as shown in figure 1. At every time step t, the agent is provided with a representation of the environment's state S_t and uses it to select an action A_t. After a time step, the agent receives a reward signal R_{t+1} and observes the system's new state S_{t+1}. The reward, should be scalar, either positive or negative, representing a bonus or penalty for achieving favourable or unfavourable outcomes, respectively.

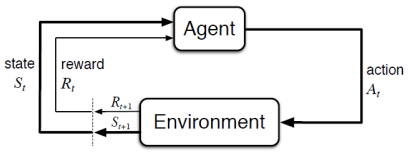

Figure 1: MDP formulation of the RL problem [19].

In the MDP framework, a mapping from every possible state $S_t = s$ to a probability of selecting every possible action $A_t = a$ is called a policy, which may be deterministic or stochastic. The return G_t of a policy is the expected cumulative reward that the policy will achieve, frequently discounted by a factor $\gamma \in [0, 1]$, as in equation 1.

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k \tag{1}$$

The definition of return allows for the introduction of the value function $V(s)$ defined in equation 2. An alternative formulation for the value function is by the definition of the expected value of the return from taking action a in state s (equation 3), called action-value function or Q-function.

$$V_\pi(s) = E_\pi[G_t|S_t = s] \tag{2}$$

$$Q_\pi(s, a) = E_\pi[G_t|S_t = s, A_t = a] \tag{3}$$

Any optimal policy $\pi^\star$ will thus be a policy that maximises the value functions $V(s)$ and $Q(s, a)$ for every state $s \in S$. From the Q-function it is also possible to generate the optimal action $a^\star = \arg\max_a Q^\star(s, a)$ directly.

Another key concept in defining the learning process for multiple RL algorithms are the Bellman equations 4 which establish a recursive definition for the value functions.

$$V_\pi(s) = E\left[r(s, a) + \gamma V_\pi(s')\right] \tag{4a}$$

$$Q_\pi(s, a) = E_\pi\left[r(s, a) + \gamma E_\pi[Q_\pi(s', a')]\right] \tag{4b}$$

2.1. Taxonomy of Reinforcement Learning Algorithms

There are multiple algorithms to approach the problem of determining the optimal policy for a given environment. Tabular methods, such as the original Q-Learning, may be only feasibly be applied in discrete and low dimensional state and action spaces, where every possible action and state may be enumerated [19].

For more complex problems, using either continuous or high-dimensional action and state spaces, using an approximator defined by parameters θ is required. The most common type of approximator are Deep Neural Networks (DNN), which have been shown to be universal non-linear approximators whose approximation power grows exponentially with the number of hidden layers, allowing them to address the "curse of dimensionality" in RL [19]. The development of a set of algorithms that use DNN-based architectures to approximate either the value function, the policy or both led to a new field of RL, Deep Reinforcement Learning (DRL).

Using differentiable approximators such as DNN, a different approach to finding the optimal policy is computing the gradient of the expected return with respect to the value function parameters and performing gradient ascent [19], a category of methods called policy optimisation or policy gradient.

A commonly used classification system for DRL algorithms is the one developed by OpenAI [1], shown in figure 2.

The first distinction presented in figure 2 is between model-based and model-free algorithms. Model-based RL requires a complete model for the

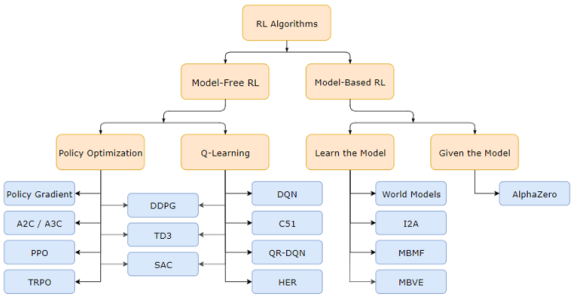

Figure 2: Taxonomy of DRL algorithms [1].

environment's transition dynamics and has higher sample efficiency [19, 1], while model-free RL learns only from information gathered from interaction with the environment. Two main frameworks are used in model-free RL: policy based methods ("Policy Optimization" in figure 2) and value-based methods ("Q-Learning" in figure 2). Current state-of-the-art methods combine the two approaches, leading to actor-critic algorithms that approximate both the policy and the value function.

In the case of the Mutriku WEC, the transition dynamics of the environment are not fully available, so model free DRL is the more natural choice for the controller. Most of the relevant variables of the model are also physical quantities that vary continuously so it is beneficial to use a DRL algorithm that is able to include continuous states and actions. This led to the choice of the three state-of-the-art DRL actor-critic algorithms in figure 2: Deep Deterministic Policy Gradient (DDPG) [15], Twin Delayed DDPG (TD3) [8] and Soft Actor-Critic (SAC) [10]. In DDPG and TD3 the policy is deterministic, so exploration noise must be added to the policy output while in SAC it is stochastic, so exploration is built in the model. DDPG is the simplest algorithm, using only periodically updated target networks to stabilise training, as well as sampling minibatches of experiences (s, a, r, s') from a replay buffer to avoid sampling consecutive highly correlated data. TD3 improves on DDPG by adding target policy smoothing noise, the clipped double-Q trick to avoid value overestimation, and periodic policy updates, while SAC introduces a policy entropy term to the loss function that promotes exploration of highly uncertain actions. All of the algorithms perform a minimisation of the MSE in estimating Bellman's equation 4b. and gradient ascent on the return from following the actor's policy simultaneously.

3. Model

A complete wave-to-wire simulation model of the Mutriku Wave Power Plant, based on previous work by Henriques et al. [12], is used as a training environment for the agent. Figure 3 describes the model by splitting it into each of its subsystems.

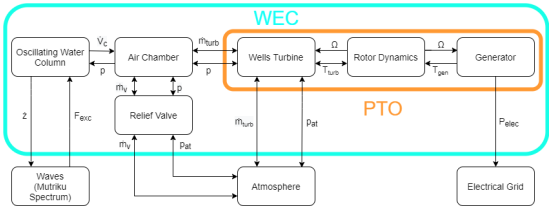

Figure 3: Mutriku model diagram.

3.1. Wave Excitation Force Generation

The local Mutriku wave climate is represented by 14 sea states SS, defined by their significant wave height, H_s, energy period, T_e, and probability of occurrence p_o, as defined in table 1 [20]. The remaining states are considered to be unable to generate significant power.

Table 1: Characteristic sea states at Mutriku [20].

Sea state number SS	Significant Height H_s (m)	Energy Period T_e (s)	Probability p_o (%)
1	0.88	5.5	3.23
2	1.03	6.5	3.44
3	1.04	7.5	5.08
4	1.02	8.5	6.11
5	1.08	9.5	10.73
6	1.19	10.5	9.31
7	1.29	11.5	9.52
8	1.48	12.5	7.42
9	1.81	13.5	2.75
10	2.07	14.5	2.96
11	2.59	15.5	1.34
12	2.88	16.5	0.40
13	3.16	11.5	0.27
14	3.20	12.5	0.42

The spectral model for the waves is thus generated from the characteristic sea states using the modified JONSWAP spectrum in equation 5 [12], where ω is the wave frequency, S_J is the original JONSWAP spectrum (a function of T_e and H_s) and $\varphi_{Mutriku}$ is a local attenuation function derived from experimental data recovered in the Mutriku site, shown in figure 4.

$$S_{Mutriku}(\omega) = S_J(\omega)\varphi_{Mutriku}(\omega) \tag{5}$$

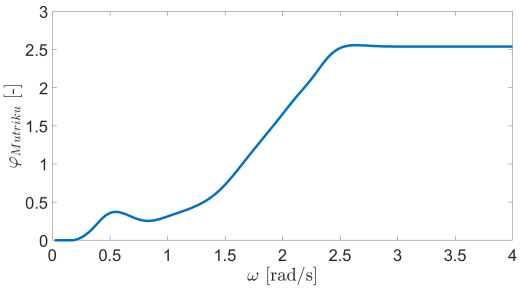

Figure 4: Attenuation function $\varphi_{Mutriku}$ [12].

The excitation force may be computed through equation 6, where Γ is the heave excitation response, ϕ is the excitation response to the wave

component, ϕ_r is a uniform random variable, and A is the amplitude of each frequency component, as computed by equation 7. The amplitudes are discretised in randomised frequency intervals $\Delta\omega_i$ as indicated by Henriques et al. [11]. Functions $\Gamma(\omega)$ and $\phi(\omega)$ are shown in figure 5.

$$F_{exc} = \sum_{i=1}^{n} \Gamma(\omega_i) A(\omega_i) \cos(\omega_i t + \phi_i(\omega) + \phi_{r,i}) \quad (6)$$

$$A(\omega_i) = \sqrt{2\Delta\omega_i S_{Mutriku}(\omega_i)} \quad (7)$$

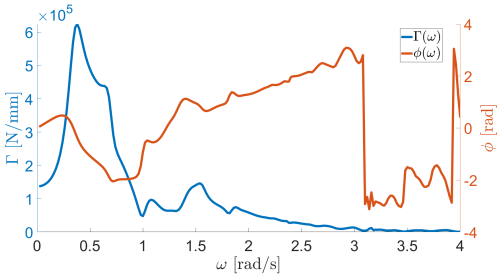

Figure 5: Response functions Γ and ϕ [12].

3.2. Water Column Hydrodynamic Model

As mentioned by Henriques et al. [12], differential equation 8 for the surface height z as a function of excitation force F_{exc} may be formulated using theory describing the motion of floating bodies, where m is the mass of the piston, A^∞ is the added mass at infinite frequency, ρ_w is the water density, g is the acceleration of gravity, S is the surface area of the OWC, p_{at} is the atmospheric pressure, $p^* = p/p_{at} - 1$ is the dimensionless air pressure in the chamber and R is the wave radiation memory term.

$$(m + A^\infty)\ddot{z} = -\rho_w g S z - p_{at} S p^* + F_{exc} - R \quad (8)$$

The values of the geometric and physical constants used in equation 8 are shown in table 2.

Table 2: OWC model Parameters.

Parameter Symbol	Value
m	72 010 kg
A^∞	27 748 kg
ρ_w	1025 kg m^{-3}
g	9.81 m s^{-2}
S	19.35 m^2
p_{at}	1.013 25 × 10^5 Pa

In equation 8, R is the memory term of the wave radiation force, expressed by the non-causal convolution integral in equation 9. This integral requires past and future wave data to be computed, so it estimated using the Prony method instead, where it is approximated by a state-space method with a set of 16 state variables I_k, derived from the representation of kernel K as a summation of complex exponential terms [4].

$$R = \int_0^t K(t-\tau)\dot{z}(\tau)d\tau \quad (9)$$

3.3. Air Chamber Expansion Model

To model the compression of the air chamber, it is common to assume the perfect gas model for air and isentropic compression [5]. It was shown [5, 12] that, under these assumptions, the differential equation that determines the dimensionless air pressure p^* is given by equation 10. In this equation, $\dot{m}$ is the mass of air flowing out of the chamber, ρ_c is the air density inside the chamber and V_c is the instantaneous air volume in the chamber, given by $V_c = V_0 - Sz$, ρ_{at} is the atmospheric air density, and $\gamma = c_p/c_v$ is the specific heat ratio of air, with a value of 1.4. At hydrostatic conditions, the air chamber has a height of 7.45 m and area S, yielding a reference volume of $V_0 = 144.1575\,\text{m}^3$.

$$\dot{p^*} = -\gamma(p^*+1)\frac{\dot{V_c}}{V_c} - \gamma(p^*+1)^{\frac{\gamma-1}{\gamma}}\frac{\dot{m}}{\rho_{at}V_c} \quad (10)$$

3.4. Turbine Dynamics Model

A set of dimensionless numbers may be defined to represent the Wells turbine dynamics, as shown in equations 11a to 11d: pressure head Ψ, flow rate Φ, power coefficient Π, and turbine efficiency η_{turb}. Additional variables used in the adimensionalisation are the turbine diameter D (0.75 m) and rotational velocity Ω, the stagnation pressure head $\Delta p = p_{at}p^*$, the inlet air density ρ_{in} and the mass flow rate $\dot{m}_{turb}$.

$$\Psi = \frac{\Delta p}{\rho_{in}\Omega^2 D^2} \quad (11a)$$

$$\Phi = \frac{\dot{m}_{turb}}{\rho_{in}\Omega D^3} \quad (11b)$$

$$\Pi = \frac{P_{turb}}{\rho_{in}\Omega^3 D^5} \quad (11c)$$

$$\eta_{turb} = \frac{P_{turb}}{P_{pneu}} = \frac{\Pi}{\Phi\Psi} \quad (11d)$$

The direction of air flow varies depending on air pressure inside the OWC chamber, changing the definition of inlet stagnation air density ρ_{in}, as shown in equation 12.

$$\rho_{in} = \begin{cases} \rho_c, & \text{if } p^* > 0 \\ \rho_{at}, & \text{if } p^* \leq 0 \end{cases} \quad (12)$$

The previously introduced dimensionless numbers were implemented in the model as functions of the dimensionless pressure head $\Phi = f_\Phi(\Psi)$, $\Pi = f_\Pi(\Psi)$ and $\eta_{turb} = f_\eta(\Psi)$, as shown in figure 6.

3.5. Generator model

Equation 13 describes the relation between the applied turbine T_{turb} and generator T_{gen} torques and the rotor's rotation velocity Ω, where I is the rotor inertia of 3.06 kg m^2.

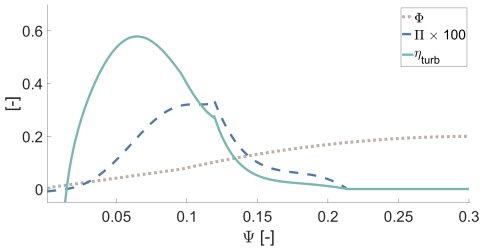

Figure 6: Dimensionless variables Φ, Π and η, as a function of Ψ [12].

$$\dot{\Omega} = \frac{T_{turb} - T_{gen}}{I} \tag{13}$$

In this model is assumed that the control braking torque will be adjusted by the generator's power electronics, and that the generator torque T_{gen} will be a resistive torque, meaning it will take a positive value in equation 13. Other restrictions that apply to the torque are the maximum generator power output $P_{gen}^{rated} = 18.5\,\text{kW}$ and the maximum generator torque $T_{gen}^{rated} = 90.1875\,\text{N m}$, meaning that under a torque control law T_{gen}^{u}, the true generator torque will be given by equation 14.

$$T_{gen} = \max\left(0, \min\left(T_{gen}^{u}, T_{gen}^{rated}, P_{gen}^{rated}/\Omega\right)\right) \tag{14}$$

The electrical power P_{elec} output from the generator may be approximated by equation 15, where η_{gen} is the generator efficiency and Λ is the generator's load factor ($\Lambda = P_{gen}/P_{rated}$) .

Previous experimental testing [9] was performed to determine the performance curve $\eta_{gen}(\Lambda)$ on a similar generator to the one used at Mutriku, demonstrating that the curve shown in figure 7 fits the experimental data [12].

$$P_{elec} = \eta_{gen}(\Lambda) P_{gen} \tag{15}$$

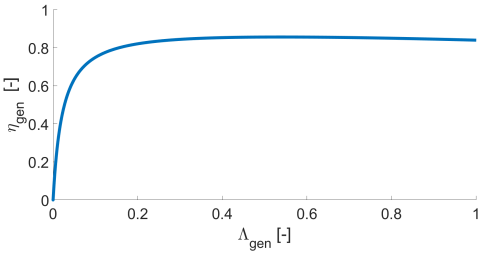

Figure 7: Generator efficiency curve η_{gen} [12].

3.6. Valve Models

Two types of valves are used in the control of OWC devices: High-Speed Safety Valves (HSSV) and relief valves, installed in series or in parallel with the turbine duct, respectively [12]. In this work, the use of a relief valve in the Mutriku OWC to control the turbine will be analysed, while the HSSV is used only as a fail-safe mechanism to prevent excessive rotation velocity.

The model for the air flow through the relief valve $\dot{m}_v$ [5], is represented by equation 16, where A_v is the effective valve area and k_v is the valve aperture state.

$$\dot{m}_v = \text{sign}(p^*) A_v k_v \sqrt{2\rho_{in}|p^* p_{at}|} \tag{16}$$

The HSSV valve is modelled as a binary variable u that is introduced in the dimensionless power function $f_\Pi(u\Psi)$, closing when the rotor reaches the maximum allowed rotation velocity $\Omega_{max} = 4000$ rpm [12].

4. Problem Definition and Controller Design

In the MDP framework used to formulate an RL problem, the environment will be the entire Mutriku simulation model described in section 3.

The observations that form the state space are the estimates for significant wave height H_s and energy period T_e from the sea state estimator, the instant dimensionless pressure p^*, the turbine rotation velocity Ω and its derivative $\dot{\Omega}$, and the generator torque T_{gen}^{t-1} and valve aperture k_v^{t-1} from the previous time step and their time derivatives $\dot{T}_{gen}^{t-1}$ and $\dot{k}_v^{t-1}$.

The action space a is formed by the time derivatives of the generator torque $\dot{T}_{gen}^{t}$ and the valve aperture $\dot{k}_v^t$. Having the controller impose a time derivative on the control actions reduces the oscillations in the control action during training, reducing the frequency of forced simulation terminations and improving convergence speed. To summarise, the action space a and state space s, are formulated in equations 17 and 18, respectively.

$$s = \begin{bmatrix} H_s & T_e & p^* & \Omega & \dot{\Omega} & T_{gen}^{t-1} & k_v^{t-1} & \dot{T}_{gen}^{t-1} & \dot{k}_v^{t-1} \end{bmatrix} \tag{17}$$

$$a = \begin{bmatrix} \dot{T}_{gen}^{t} & \dot{k}_v^t \end{bmatrix} \tag{18}$$

The reward function r_t must be defined according to the control objectives: maximise power production, avoid excessive PTO control effort and preserve the structural integrity of the system.These requirementes lead to equation 19.

In this equation, a positive reward is given proportional to the average normalised electrical power $\overline{P_{elec}}$, taken as the average of the output power over the controller's sampling interval, normalised to the $[0, 1]$ range by the maximum possible electric power output by the generator. Negative terms are added proportional to the dimensionless generator torque and valve aperture in the previous time step $T_{gen}^{t-1}/T_{gen}^{rated}$ and k_v^{t-1} to discourage excessive control effort. A penalty is also added in proportion to the time derivative of generated power raised an even power e_{even}, to penalise large variations in the output power.

Proportionality constants k_1, k_2, k_3 and k_4 allow for separate tuning of the importance of each of

the reward terms. Flag f_1 takes a value of 1 if the simulation is terminated early due to turbine stoppage or due to the dimensionless pressure in the chamber reaching a threshold value of $|p^*| > 0.25$. If the controller outputs a torque that requires a a higher generator output power than the rated power ($P_{gen} > P_{gen}^{rated}$), flag f_2 is activated. Finally, if the generator-turbine set reaches its maximum rotation velocity Ω_{max}, flag f_3 will activate.

$$\begin{aligned} r_t =& k_1 \overline{P_{elec}} - k_2 \left(\frac{d}{dt} \overline{P_{elec}} \right)^{e_{even}} \\ & - k_3 \frac{T_{gen}^{t-1}}{T_{gen}^{rated}} - k_4 k_v^{t-1} - \sum_{i=1}^{3} f_i \end{aligned} \tag{19}$$

The values of the proportionality constants $k_{1,4}$ and flags $f_{1,3}$ are given in table 3.

Table 3: Reward constants and flags.

k_1	k_2	k_3	k_4	f_1	f_2	f_3	e_{even}
10	4	0.1	0.05	−20	−5	−0.5	8

4.1. Sea State Estimation

The JONSWAP spectrum may be used directly in simulation to generate the excitation force from a sea state, but in the context of online controller deployment, H_s and T_e must be estimated from time series data that is representative of the wave dynamics. In this implementation, one of the challenges was using exclusively air chamber pressure data to approximate the sea spectrum $\hat{S}(\omega)$.

Spectral analysis may be used to characterise sea states since, given a power spectral density $S(\omega)$, wave significant height H_s and energy period T_e are defined by equations 20 and 21 [17].

$$H_s = 4\sqrt{\int_0^\infty S(\omega) d\omega} \tag{20}$$

$$T_e = \frac{\int_0^\infty \frac{S(\omega)}{\omega} d\omega}{\int_0^\infty S(\omega) d\omega} \tag{21}$$

To approximate the power spectral density of the pressure time series, the modified periodogram is used [18], which is given by equation 22.

$$\hat{S}(f) = \frac{\Delta_t}{N} \left| \sum_{t=1}^{N} w(t) y(t) e^{-i 2\pi f \Delta_t t} \right|^2 \tag{22}$$

After an extensive parameter search, the values for the periodogram parameters were selected. These include the sampling interval $\Delta_t = 1\,\text{s}$, the number $N = 900$ of time series points to use, the window function $w(t) = 1$ (unmodified periodogram) and the update interval $t_{spectrum} = 5\,\text{s}$.

To validate the algorithm, online estimation was performed for a 15 min period on each sea state. Figure 8 shows the results of this simulation. While the reconstruction of the sea states is not perfect, the qualitative ordering of the values of T_e and H_s is preserved in general, with only temporary errors caused by random pressure peaks. The most likely cause for estimation errors is the non-linearity of the model, which includes the air compression, the frequency dependent attenuation function $\varphi_{Mutriku}$ and the limit on pressure due to the safety valve.

Another relevant issue is the oscillations in the estimate of the energy period (figure 8), which stem from numerical issues in the quotient of integrals in equation 21, minimised by only recomputing the estimate every $t_{spectrum} = 5\,\text{s}$.

4.2. Controller Architecture

The chosen DRL methods require choosing a set training hyperparameters, including the DNN structure and a set of numerical parameters that control the learning process.

Since TD3 is an extension to DDPG that improves training stability, to properly compare these two alternatives it was decided to use the same actor and critic network architectures for both, introducing only a second critic in the TD3 agent. The chosen architectures for the actor and critic networks are shown in the left and middle networks in figure 9, respectively. The network is mostly made up of a series of fully connected (FC) and rectified linear unit (ReLU) layers. Note the parallel input paths for state and action in the critic, and the tangent layer as the actor output layer, used to squash the output to the [-1,1] interval.

To increase training stability, the size of the replay buffer and mini-batch were increased from their default implementations [15, 8] to 100000 and 256, respectively. To avoid over-fitting, a gradient threshold of 1 was imposed on the gradients used in optimization and an L2 regularisation term was added to the loss function. All other hyperparameters were kept from the original implementations.

Unlike the DDPG and TD3 controllers, the SAC controller uses a stochastic actor representation, meaning that while the structure for the pair of critic networks may be reused from TD3 (middle network in figure 9), the actor network must be modified to output a normal distribution, as shown in the right network in figure 9. This is achieved by splitting the network output path into two branches representing the mean and standard deviation of the stochastic policy. Since the standard deviation must take a smooth positive value, a softplus layer is added on the corresponding branch before the output.

The numeric hyperparameters for SAC were either reused from DDPG and TD3 in order to allow for a direct comparison of the algorithms, or, for the ones that are unique to SAC, the values presented in the original paper [10] were used

5. Results

Each of the controllers was trained using the MDP formulation and training hyperparameters found in

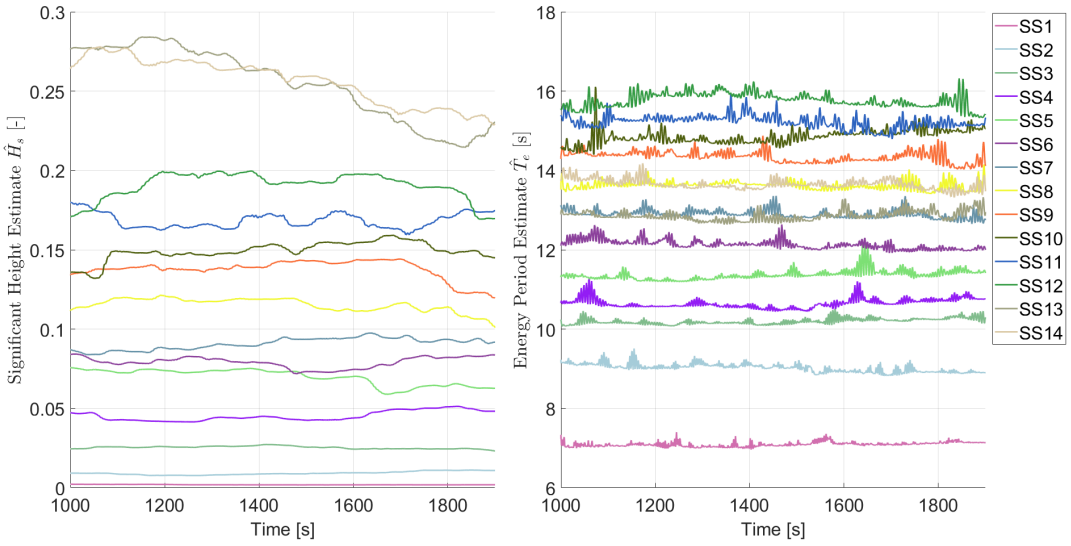

Figure 8: Real time estimates for the significant height and energy period for each of the sea states.

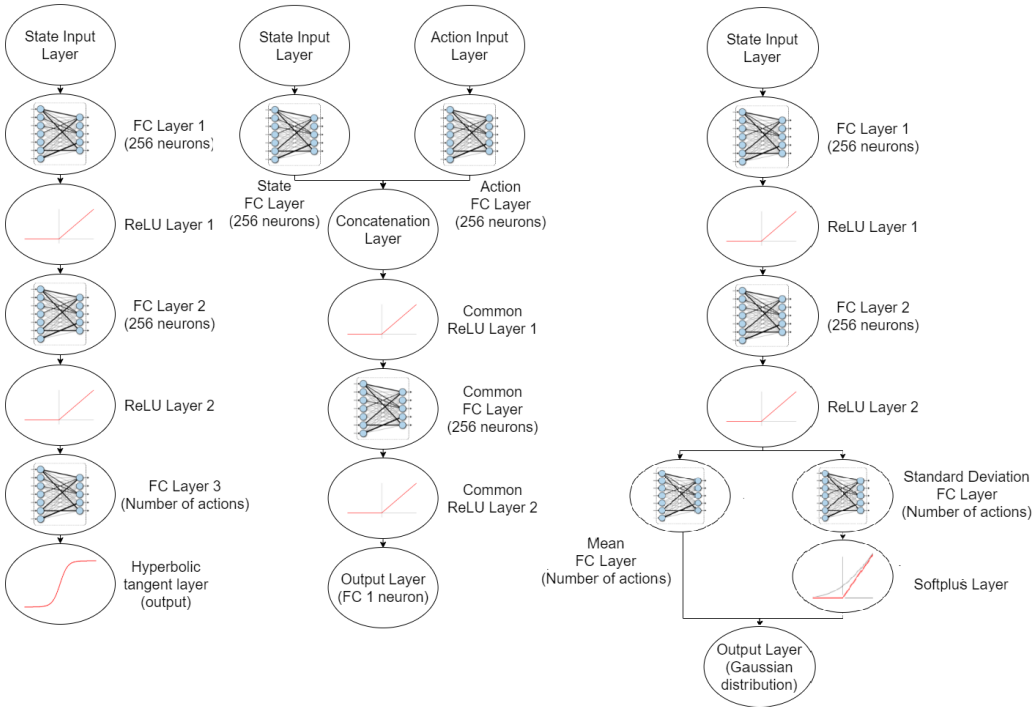

Figure 9: Left to right: DDPG and TD3 actor, critic and SAC actor DNN architectures.

the previous section. Since the possible reward from an episode is largely dependent on the randomly sampled sea state, it is inconvenient to set an average reward target as a stopping criterion, so training is stopped at a set number of simulation time steps, with convergence evaluated after training.

In order to expose the agents to all of the characteristic sea states, every episode is initialised by choosing a random sea state, and simulated 30 min of simulated time in normal operation, unless early termination is forced by flag f_1.

5.1. Training Process

The training curve for DDPG and TD3 is shown in figure 10. Comparing the three controllers, it is clear that the DDPG controller converges to a policy that achieves a lower average reward than TD3 or SAC, showing the advantage of introducing the two latter, more complex algorithms. The critic output evolution also shows value overestimation in DDPG, where the critic estimates a higher average value than TD3 or SAC while achieving a lower average reward. DDPG was also prone to over-fitting and divergence, requiring early stopping of training at 1×10^5 steps, instead of the 3×10^5 used in TD3 and SAC, to obtain a viable controller.

The training time, simulated time and ratio between them for each of the trained controllers is shown in table 4. Every controller presents a ratio higher than 1, showing the viability of implementing training in real time on a physical prototype,

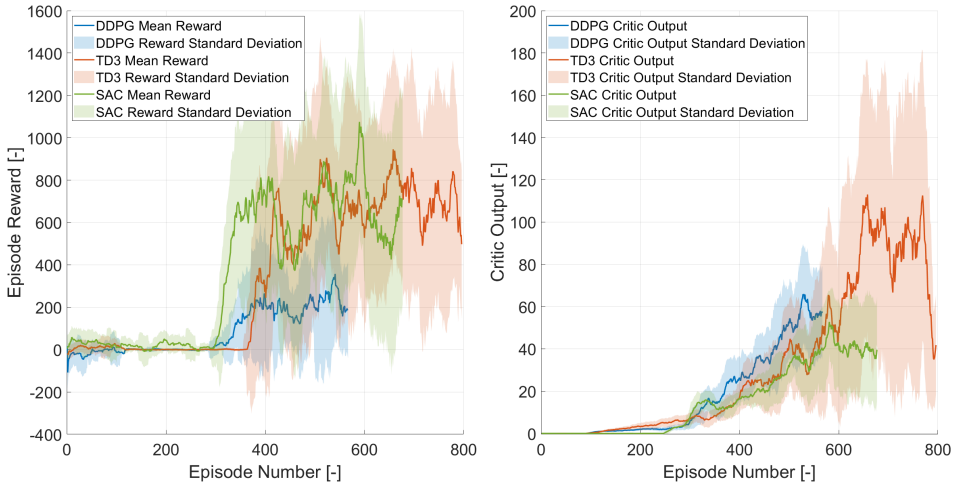

Figure 10: DDPG, TD3 and SAC controllers training reward curves.

even with consumer level hardware.

Table 4: Training and simulation time for the 3 DRL algorithms.

DRL algorithm algorithm	Training time [s]	Simulated time [s]	Time ratio
DDPG	25 443	200 444	7.87
TD3	76 664	600 132	7.82
SAC	117 970	601 682	5.10

5.2. Controller performance

To evaluate controller performance, the average power generation is calculated over 20 random initialisations of each sea state simulated for 30 min. The controllers are compared not only against each other but with the optimal baseline law $T_{gen}(\Omega) = 2 \times 10^4 \Omega^2$ [12], yielding the results in figure 11.

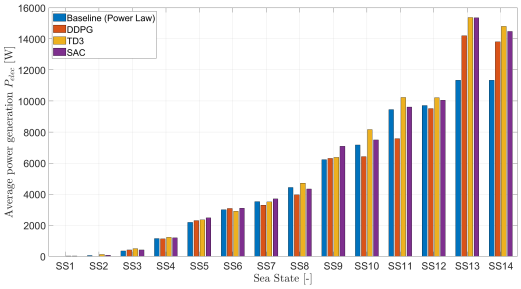

Figure 11: Electrical power generation under each control law.

The first conclusion to be drawn is that the DDPG agent is not an adequate solution to this problem, since it only outperforms the baseline power law on sea states SS3 to SS5, SS9, SS13 and SS14. Comparing it to the other two DRL controllers, it only outperforms TD3 or SAC on sea state SS6 and SS3, respectively.

TD3 and SAC present similar power generation values for most sea states, but TD3 has a higher mean power generation on the more (SS10 to SS14 and SS8) and less (SS1 to SS4) energetic sea states, while SAC favours electricity production in states with an intermediate significant height (SS5 to SS7 and SS9).

For the intermediate sea states (SS3 to SS12, the TD3 and SAC controller are able to achieve slightly higher power generation on average than the power law controller, with increases varying from 1% to 16% over the baseline.

The improvement seen in power generation is larger for the most energetic sea states, particularly SS13 and SS14 with improvements of 35.4% and 27.7%, respectively, when using SAC or 35.6% and 31.1% when using TD3.

Low energy sea states also benefit from a significant increase in power generation under SAC and TD3 over the baseline law, which was expected as the power law optimises for turbine efficiency, ignoring generator efficiency which is lower when the generator operates at a smaller load factor. By including this effect, power generation is increased by 112%, 136% and 39% in sea states SS1, SS2 and SS3, respectively, for the TD3 agent, and by 100%, 43.5% and 16.8% for the SAC agent, when compared with the baseline values.

Using the probability distribution p_o from table 1 over a year, the expected value of the electrical power generation may be estimated. Furthermore, considering the average Iberian Electricity Market (MIBEL) price in September 2021 as a reference (160.77 €/MWh) [16], the expected power generation and revenue under each control law are shown in table 5, assuming operation of all 14 Wells turbines at Mutriku over a year [12].

Table 5 verifies that DDPG does not represent a viable alternative for this problem, since it is not expected to generate higher yearly energy than the baseline, leading to an expected yearly loss for the plant operator of 1273€. TD3 and SAC, however, present a clear advantage over the baseline law, yielding additional profits of 2227€ and 2285€, respectively. The SAC expected power generation is

able to surpass TD3 over a full operation year due to its better performance on the most likely sea states to occur (SS5 to SS7).

Table 5: Yearly nergy generation and revenue under each control scheme.

Control algorithm	Energy Generation [MWh]	Total Revenue [€]
Power Law	232.8	37423
DDPG	224.9	36150
TD3	246.6	39650
SAC	247.0	39708

Analysing the behaviour of the two successful control schemes, figures12 and 13 show a set of relevant OWC variables when applying the trained SAC and TD3 controllers, respectively, on representative sea states 3, 8 and 13.

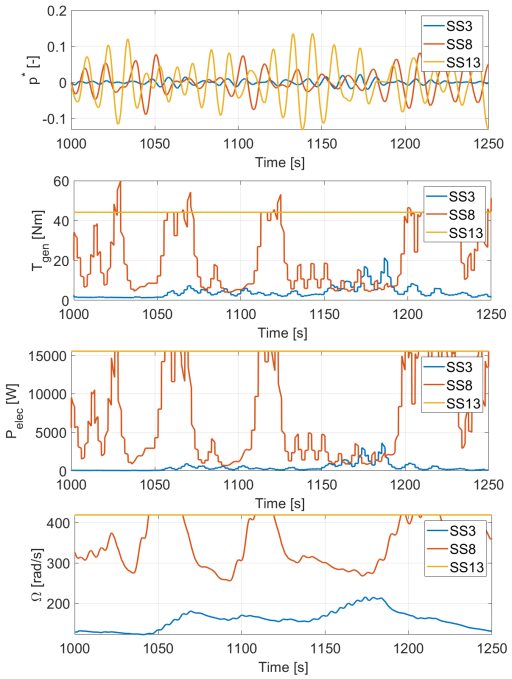

Figure 12: Plant performance using the SAC controller on sea states 3, 8 and 13.

Compared with the power law controller, the most notable change is the presence of sharp peaks in the generator torque for sea states 3 and 8. This behaviour may be explained in part by considering the optimality conditions described in frequency domain control, well-approximated by a type of latching control. This behaviour is even further encouraged when considering the generator efficiency curve (figure 7), which shows that efficiency is higher at higher generator loads. Unlike the power control law, the agent includes the generator efficiency in its behaviour so, in sea states where the available pneumatic power is low, it lowers the electromagnetic torque to allow rotation velocity to increase, and then extracts the kinetic energy stored by inertia at a higher load factor by suddenly increasing the applied torque.

A similar behaviour occurs when using TD3 to control the plant, where an improvement compared to SAC is seen mainly in SS8, where the controller manages to keep the rotation velocity mostly under the reference value of 4000 rpm, unlike the SAC controller where the HSSV must be actuated in regular intervals.

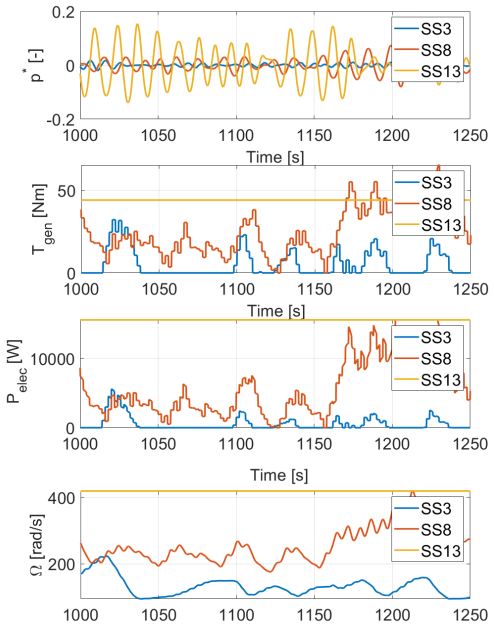

Figure 13: Plant performance using the TD3 controller on sea states 3, 8 and 13.

6. Conclusions

In this work, three for DRL agent architectures were applied to the problem of maximising electrical power generation in the Mutriku OWC.

Using this type of algorithm to control the plant shifts the main computational effort to the training process, that may occur offline and in simulation, before deployment on a prototype. However, this does not exclude the possibility of performing online training as well, allowing the controller to adapt to changing system dynamics. DRL also has the advantage of being a model-free technique, making it a useful approach when facing systems with high modelling uncertainty, as is the case of the Mutriku OWC, as well as being able to retrain to any other OWC device with minimal adaptation.

Analysing the training results, the DDPG controller has poor performance, due to instability issues and convergence to a sub-optimal policy, which resulted in a lower expected power production than the baseline power law. In contrast, TD3 and SAC have both been shown to be promising alternatives, with the trained controllers leading do an increase in expected yearly electric power production of 5.9% and 6.0% over the baseline.

Future research on the application of Deep Reinforcement Learning to the control of OWC devices should focus on the testing of these types of algorithm on a physical prototype. To ensure a safe

training process, training should begin in simulation and then be refined by transferring the controller to the real system for further training

It would also be beneficial for controller robustness to use real pressure data from the Mutriku site as an input to the simulation, but this approach may not be valid in cases where the turbine operation has an effect on the chamber pressure.

Acknowledgements
The author would like to thank Professors Miguel Ayala Botto and Susana Vieira for supervising this work, and Professor João Henriques and Engineer Jorge Silva for providing the Mutriku model and its parameters.

References

[1] J. Achiam. Spinning Up Documentation. OpenAI, 2020. Available at `https://spinningup.openai.com/` (accessed: 25-05-2021).

[2] E. Anderlini, D. I. M. Forehand, P. Stansell, Q. Xiao, and M. Abusara. Control of a Point Absorber Using Reinforcement Learning. *IEEE Transactions on Sustainable Energy*, 7(4):1681–1690, Oct. 2016.

[3] E. Anderlini, S. Husain, G. G. Parker, M. Abusara, and G. Thomas. Towards Real-Time Reinforcement Learning Control of a Wave Energy Converter. *Journal of Marine Science and Engineering*, 8(11), Oct. 2020.

[4] G. Duclos, A. H. Clément, and G. Chatry. Absorption of outgoing waves in a numerical wave tank using a self-adaptive boundary condition. *International Journal of Offshore and Polar Engineering*, 11(3):168–175, 2001.

[5] A. O. Falcão and P. Justino. OWC wave energy devices with air flow control. *Ocean Engineering*, 26(12):1275–1295, Dec. 1999.

[6] J. Falnes. Optimum control of oscillation of wave-energy converters. *International Journal of Offshore and Polar Engineering*, 12(2):147–155, 2002.

[7] F. X. Faÿ, J. C. Henriques, J. Kelly, M. Mueller, M. Abusara, W. Sheng, and M. Marcos. Comparative assessment of control strategies for the biradial turbine in the Mutriku OWC plant. *Renewable Energy*, 146:2766–2784, 2020.

[8] S. Fujimoto, H. van Hoof, and D. Meger. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 4, pages 2587–2601, Stockholm, Sweden, July 2018.

[9] L. Gato, A. Carrelhas, F. C. da Fonseca, and J. Henriques. Opera deliverable d3.2 - turbine-generator set laboratory tests in variable unidirectional flow. Technical report, Instituto Superior Técnico, 2017.

[10] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 5, pages 2976–2989, Stockholm, Sweden, July 2018.

[11] J. Henriques, M. Lopes, R. Gomes, L. Gato, and A. Falcão. On the annual wave energy absorption by two-body heaving WECs with latching control. *Renewable Energy*, 45:31–40, Sept. 2012.

[12] J. Henriques, J. Portillo, W. Sheng, L. Gato, and A. Falcão. Dynamics and control of air turbines in oscillating-water-column wave energy converters: Analyses and case study. *Renewable and Sustainable Energy Reviews*, 112:571–589, Sept. 2019.

[13] International Energy Agency (IEA). *Electricity Information Overview*. International Energy Agency, 2020.

[14] IRENA. *Innovation Outlook: Ocean Energy Technologies*. International Renewable Energy Agency, Abu Dhabi, 2020.

[15] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, San Juan, Puerto Rico, May 2016.

[16] OMIE. Annual Final Energy, 2021. Available at `www.omie.es/en/market-results/` (accessed: 27-10-2021).

[17] W. Sheng and H. Li. A Method for Energy and Resource Assessment of Waves in Finite Water Depths. *Energies*, 10(4):460, Apr. 2017.

[18] P. Stoica and R. L. Moses. *Spectral Analysis of Signals*. Pearson Prentice Hall, 1^{st} edition, 2005.

[19] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Massachussets, 2^{nd} edition, 2018.

[20] Y. Torre-Enciso, I. Ortubia, L. I. López de Aguileta, and J. Marqués. Mutriku Wave Power Plant: from the thinking out to the reality. In *8th European Wave and Tidal Energy Conference (EWTEC 2009)*, pages 319–328, Uppsala, Sweden, 2009.