Control of wave energy converters using machine learning strategies

Enrico Anderlini

Engineering Doctorate

Industrial Doctoral Centre for Offshore Renewable Energy

2017

IDCORE

This thesis is submitted in partial fulfilment of the requirements for the award of an Engineering Doctorate, jointly awarded by the University of Edinburgh, the University of Exeter, and the University of Strathclyde.
The work presented has been conducted under the industrial supervision of Wave Energy Scotland as a project within the Industrial Doctoral Centre for Offshore Renewable Energy (IDCORE).

Abstract

Wave energy converters are devices that are designed to extract power from ocean waves. Existing wave energy converter technologies are not financially viable yet. Control systems have been identified as one of the areas that can contribute the most towards the increase in energy absorption and reduction of loads acting on the structure, whilst incurring only minimal extra hardware costs. In this thesis, control schemes are developed for wave energy converters, with the focus on single isolated devices.

Numerical models of increasing complexity are developed for the simulation of a point absorber, which is a type of wave energy converter with small dimensions with respect to the dominating wave length. After investigating state-of-the-art control schemes, the existing control strategies reported in the literature have been found to rely on the model of the system dynamics to determine the optimal control action. This is despite the fact that modelling errors can negatively affect the performance of the device, particularly in highly energetic waves when non-linear effects become more significant. Furthermore, the controller should be adaptive so that changes in the system dynamics, e.g. due to marine growth or non-critical subsystem failure, are accounted for. Hence, machine learning approaches have been investigated as an alternative, with a focus on neural networks and reinforcement learning for control applications. A time-averaged approach will be employed for the development of the control schemes to enable a practical implementation on WECs based on the standard in the industry at the moment.

Neural networks are applied to the active control of a point absorber. They are used mainly for system identification, where the mean power is related to the current sea state and parameters of the power take-off unit. The developed control scheme presents a similar performance to optimal active control for the analysed simulations, which rely on linear hydrodynamics.

Reinforcement learning is then applied to the passive and active control of a wave energy converter for the first time. The successful development of different control schemes is described in detail, focusing on the encountered challenges in the selection of states, actions and reward function. The performance of reinforcement learning is assessed against state-of-the-art control strategies. Reinforcement learning is shown to learn the optimal behaviour in a reasonable time frame, whilst recognizing each sea state without reliance on any models of the system dynamics. Additionally, the strategy is able to deal with model non-linearities. Furthermore, it is shown that the control scheme is able to adapt to changes in the device dynamics, as for instance due to marine growth.

Acknowledgements

First of all, I would like to thank my academic supervisors Doctors Mohammad Abusara and David Forehand for their continuous support and advice during my project. Without their help, I would not have been able to overcome some of the challenges caused by unexpected circumstances. I would also like to thank my industrial supervisor Elva Bannon for her support. I want to recognize the help I got from Doctors David Pizer and Paul Stansell when I was at Pelamis Wave Power, from whom I have learned many a thing on wave energy and reinforcement learning. Similarly, I would like to thank Professor Gordon Parker from Michigan Technological University for his help with programming and control systems during his 6-month-long stay at the University of Edinburgh. I would also like to recognize the collaboration I have had with Anup Nambiar at the University of Edinburgh on adapting reinforcement learning for a tidal energy application. My gratitude goes also to Doctor Qing Xiao at the University of Strathclyde for providing guidance on the original project on the estimation of wake losses on wind turbines and for organizing the viva voce.

In addition, I would like to thank all people involved with delivering the IDCORE program, which I have enjoyed very much. Particular thanks should go to Ksenia Siedlecka, who aided me to find a project with Pelamis Wave Power after the folding of the original project on wind energy by Lloyd's Register. Professor David Ingram and lawyer Lindsay Hampton helped me move my current project to Wave Energy Scotland after the bankruptcy of Pelamis Wave Power. More recently, Stephen Graham has provided me guidance with the completion of the tasks required by the IDCORE programme.

Finally, I would like to thank all students within the IDCORE programme for the enjoyable four years I have had. I would also like to recognize the fellow researchers at the Institute of Energy Systems at the University of Edinburgh with whom I have shared an office space, lunches and nights out.

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Enrico Anderlini

Date: 2nd November 2017

List of Symbols

As this thesis deals with the topics of wave energy, neural networks and reinforcement learning, many symbols are used, which may take on different meanings depending on the context. As a result, here the reader can find a list of symbols which should make the thesis easier to read.

Linear wave theory and WEC modelling

Note that the hat indicates complex numbers (complex notation is useful in the frequency domain). An arrow indicates three-dimensional vectors, while bold small letters indicate vectors of higher order. Matrices are indicated with bold capital letters.

x	displacement in surge
y	displacement in sway
z	displacement in heave
ϕ	displacement in roll
θ	displacement in pitch
ψ	displacement in yaw
$\vec{v}$	flow velocity vector
$\vec{f}$	vector of gravitational force per unit volume
p	fluid pressure
μ	fluid viscosity
Φ	velocity potential
g	gravitational acceleration
C	integration constant
$\vec{n}$	unit normal vector
$\vec{v_\mathrm{b}}$	body velocity vector
h	water depth
ζ	wave elevation
ω	circular wave frequency
β	wave direction angle
k	wave number
λ	wave length
ζ_a	wave amplitude
p_atm	atmospheric pressure
ρ	water density
$\hat{p}$	complex hydrodynamic pressure
v_p	propagation velocity
v_g	wave group velocity
$\vec{f}$	three-dimensional hydrodynamic force vector
$\vec{m}$	three-dimensional hydrodynamic moment vector
$\boldsymbol{f}$	six-dimensional generalised force vector
$\vec{s}$	position of infinitesimal surface element
$\vec{u}$	three-dimensional linear velocity vector
$\vec{\Omega}$	three-dimensional angular velocity vector
$\boldsymbol{n}$	six-dimensional normal vector
$\boldsymbol{v}$	six-dimensional generalised velocity vector
Φ_i	incidence velocity potential
Φ_d	diffraction velocity potential
Φ_r	radiation velocity potential
$\boldsymbol{\phi}$	radiated velocity potential vector
R	distance from rigid body
S	body wetted surface area

$\boldsymbol{f}_{\mathrm{r}}$	radiated force vector (either frequency- or time-domain depending on context)
$\boldsymbol{Z}$	radiation impedance matrix
$\boldsymbol{A}$	added mass matrix
$\boldsymbol{B}$	hydrodynamic damping matrix
$\boldsymbol{f}_{\mathrm{e}}$	excitation force vector (either frequency- or time-domain depending on context)
$\boldsymbol{H}$	vector of incidence and diffraction velocity potentials per wave amplitude
$\boldsymbol{\xi}$	six-dimensional position vector in an inertial reference frame
$\boldsymbol{\xi}_0$	six-dimensional equilibrium position
$\boldsymbol{\eta}$	six-dimensional displacement vector
$\boldsymbol{f}_{\mathrm{h}}$	hydrostatic restoring force vector (either frequency- or time-domain)
$\boldsymbol{C}$	hydrostatic restoring matrix
$\boldsymbol{C}_{\mathrm{h}}$	hydrostatic restoring matrix for multiple bodies
$\boldsymbol{M}$	mass matrix
$\boldsymbol{a}$	six-dimensional generalized acceleration vector
$\boldsymbol{I}$	identity matrix
$\boldsymbol{I}^G$	inertia tensor
G	centre of gravity
m	body mass
$\boldsymbol{v}$	six-dimensional generalised velocity vector
$\boldsymbol{K}$	radiation impulse response function matrix
$\boldsymbol{\mu}$	matrix of constants
$\boldsymbol{h}$	excitation impulse response vector
$\boldsymbol{H}$	Fourier transform of the excitation impulse response function
γ	phase angle
t	time
R	ramp function
t_{r}	duration of initial ramp function
S	wave spectrum
n_{w}	number of individual wave components
S_{B}	Bretschneider wave spectrum
S_{J}	JONSWAP wave spectrum
H_{s}	significant wave height
T_{p}	peak wave period
ω_{p}	peak circular wave frequency
γ	non-dimensional peak shape parameter
σ	JONSWAP shape function
$\Delta\omega$	circular wave frequency step
a	amplitude of each wave component

T_e	energy wave period
P	polynomial of zeros
Q	polynomial of poles
$\boldsymbol{x}_\mathrm{ss}$	state-space vector for radiation approximation
$\boldsymbol{A}_\mathrm{ss}$, $\boldsymbol{B}_\mathrm{ss}$, $\boldsymbol{C}_\mathrm{ss}$ and $\boldsymbol{D}_\mathrm{ss}$	state-space matrices
$\boldsymbol{f}_\mathrm{m}$	mooring force vector
$\boldsymbol{B}_\mathrm{m}$	mooring damping matrix
$\boldsymbol{C}_\mathrm{m}$	mooring stiffness matrix
$\boldsymbol{f}_\mathrm{d}$	viscous damping force vector
$\boldsymbol{C}_\mathrm{d}$	viscous drag matrix
A_d	characteristic area
$\boldsymbol{v}_\mathrm{G}$	generalised vector of unperturbed flow velocity
$\boldsymbol{f}_\mathrm{PTO}$	PTO force vector
$\boldsymbol{P}$ and $\boldsymbol{Q}$	matrices used in the assembly of the time domain model of a WEC
w	velocity in heave

Model of Seabased WEC

I_s	stator current
P_nom	nominal power of generator
V_nom	nominal voltage of generator
v_nom	nominal speed of generator
l_p	piston length
l_s	stator length
k_u	stiffness of upper end stop
k_l	stiffness of lower end stop
k_w	stiffness of wire
k_s	stiffness of restoring spring
m_b	mass of float
m_p	mass of piston
f_u	force of upper end stop
f_l	force of lower end stop
f_em	electromotive or PTO force
f_s	force of restoring spring

Control of WECs

B_{PTO}	PTO damping coefficient
C_{PTO}	PTO stiffness coefficient
$B_{\mathrm{PTO,opt}}$	optimal PTO damping coefficient
$C_{\mathrm{PTO,opt}}$	optimal PTO stiffness coefficient
$\dot{z}$	velocity in heave
$\bar{P}_{\mathrm{opt}}$	mean optimal power
k_{τ}	generator torque constant
i_{s}	current of stator
A_{fac}	active area
b	constant
P	absorbed power
η	PTO system efficiency
$\mathcal{L}$	capture width
P_{w}	mean power per unit crest
c_{g}	wave group velocity
D	diameter of point absorber

Neural networks

$\boldsymbol{x}$	input vector
$\boldsymbol{w}$	weight vector
d	signal from each input node
y	neuron output
b	bias term for each neuron
n	index of individual neuron
l	index of individual layer
$\boldsymbol{W}^l$	weight matrix between every two layers
$\boldsymbol{b}^l$	bias vector between every two layers
f^l	activation function of each layer
$\boldsymbol{o}^l$	output vector of each layer
J	performance index or cost function
$\boldsymbol{y}$	output vector of neural network
$\boldsymbol{y}_{\mathrm{tr}}$	input vector of training data
$\boldsymbol{y}_{\mathrm{tr}}$	output vector of training data
λ	scale of regularization term
α	learning rate
$\boldsymbol{\delta}$	sensitivity vector
$\boldsymbol{J}$	Jacobian of the vector of independent variables
μ	weighting variable

Reinforcement learning

The tilde denotes learnt values.

s	current state
s'	next state
a	current action
a'	next action
r	reward associated with current state and action
Q	state-action value function
V	state value function
$\mathcal{S}$	state variables space
$\mathcal{A}$	action space
$\mathcal{P}$	probability of transitioning to a particular state
γ	discount factor
$\mathcal{R}$	reward corresponding to a particular transition
π	policy
Ω	set of all probability distributions
E	expected reward
$\boldsymbol{P}$	matrix of the transition model
$\boldsymbol{\Pi}$	matrix of possible policies
$\boldsymbol{I}$	identity matrix
ϵ	exploration rate
ϵ_0	initial exploration rate
α	learning rate
α_0	initial learning rate
$\boldsymbol{N}$	matrix of the number of visits to each state-action pair
N_ϵ	parameter of minimum number of visits to a specific state before reducing the exploration rate
N_α	parameter of minimum number of visits to a specific state-action pair before reducing the learning rate
$\boldsymbol{R}$	list of all returns
$\boldsymbol{\phi}$	vector of features
$\boldsymbol{\Theta}$	weight matrix
$\boldsymbol{\mu}$	vector of bandwidth of radial basis functions
$\boldsymbol{s}$	vector of positions of the centres of radial basis functions
$\boldsymbol{i}$	input vector to neural network
$\boldsymbol{o}$	output vector from neural network
e	mean squared error of neural network
Q_{target}	target Q-value
$\boldsymbol{S}$	list of samples
T_π	Bellman operator
$\boldsymbol{\Phi}$	matrix of features
$\boldsymbol{w}$	weight vector

μ	probability distribution
$\boldsymbol{\Delta}_{\mu}$	diagonal matrix of projection weights
$\boldsymbol{A}$ and $\boldsymbol{b}$	matrix and vector used in determination of weights of function approximation

WEC control based on neural networks and reinforcement learning

H	wave height of regular waves
T	wave period of regular waves
P_{avg}	averaged absorbed power
B_{min}	minimum allowable PTO damping coefficient
B_{max}	maximum allowable PTO damping coefficient
C_{min}	minimum allowable PTO stiffness coefficient
C_{max}	maximum allowable PTO stiffness coefficient
c	cost function
N_{h}	number of time horizons
h	index of time horizon
$D(h)$ or H_{RL}	time horizon duration
t_{i}	initial time of the episode
t_{f}	final time of the episode
E	energy extracted over an episode
$\boldsymbol{E}$	mean energy for each state-action combination
$\boldsymbol{N}_{\mathrm{c}}$	matrix of number of visits to each state-action combination
ΔB_{PTO}	step change in PTO damping coefficient
ΔC_{PTO}	step change in PTO stiffness coefficient
$\boldsymbol{m}$	vector of mean reward values for each discrete state
p	penalty term
δ_{c}	distance between kernels of radial basis functions
M	number of radial basis functions employed
z_{lim}	soft displacement constraint in heave

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

Economical, sustainable and secure energy sources are fundamental for continuous development of modern economies. Since the industrial revolution, fossil fuels have represented a major energy supply due to their high energy density, ease of storage and distribution and low extraction costs. Nevertheless, most countries have recently started to shift their energy policy towards more sustainable and secure energy sources. Most scientists have recognized the negative impact of green-house emissions associated with fossil fuels (Maslin, 2014). The role of renewable energy in power generation has thus been increasing substantially in importance, due to its sustainable and secure nature. In particular, solar and wind power can be considered to have reached commercial maturity. Conversely, wave energy technologies are not financially viable yet, despite the enormous resource potential of up to 2.1 TW of power worldwide (Gunn and Stock-Williams, 2012).

The main issue associated with wave energy is its high energy content, which requires strong, sturdy, complex machines that are able to withstand high loadings. This problem is exacerbated by the high peak-to-mean of wave power, which has to be addressed with devices optimized to absorb as much energy as possible in calm sea states and at the same time survive the worst storms. In addition, locations with a high energy resource tend to be located in remote areas, which increases operation and maintenance costs. Furthermore, matching supply and demand is difficult like with renewable energy sources, as the availability is dictated by environmental conditions. Nevertheless, the increase in wave energy content in winter months sits well with standard energy demand patterns in countries located north and south of the tropics, such as the United Kingdom.

In order to reach commercial maturity, the levellised cost of energy produced by wave energy converters (WECs) needs to be brought down from current levels. In particular, Wave Energy Scotland (2017) has identified the requirement to bring the levellised cost of energy below £150/MWh as fundamental. Control systems have been identified as an area that can significantly improve the financial viability of WEC technologies (Wave Energy Scotland, 2017), as they may be used to increase energy absorption, reduce loads and ensure the compliance with motion and power flow constraints.

This thesis will present the development of innovative control strategies for WECs based on methods developed by the artificial intelligence community.

1.1 Motivation

WECs convert the oscillation of kinetic and potential energy carried by ocean gravity waves to electrical energy that can be delivered to the electrical grid through a mechanism known as power take-off system (PTO). The PTO may include intermediate hydraulic and/or mechanical stages. By controlling the force that the PTO exerts on the WEC, it is possible to tune the system dynamics to the incoming waves for the maximization of energy absorption and/or reduction of loads on the structure. As aforementioned, Wave Energy Scotland (2017) has identified the development of suitable control strategies as of fundamental importance for the development of the wave energy industry, with an expected decrease in levellised cost of energy.

The schemes for WEC control can be classified into real-time and time-averaging approaches. The former strategies have become the focus of substantial research work by academics in the past decade, particularly model predictive control (Ringwood *et al.*, 2014; Korde and Ringwood, 2016). Nevertheless, the practical uncertainties related to the measurement and prediction of the incoming wave elevation profile mean that time-averaging approaches have been mostly applied on real devices by the wave energy industry to date (Wave Energy Scotland, 2016). These strategies assume stationary sea state conditions for intervals lasting 15-30 minutes, with the controller parameters optimized for the average conditions during that time. The optimized controller parameters are computed from simulations and validated through experimental testing (Wave Energy Scotland, 2016). Due to the industrial nature of the project, these methods will be investigated in this thesis.

Modelling errors in the determination of the control parameters can have a negative impact on the performance of the WEC. In particular, non-linear effects are more important in waves with a higher energy content, where damage to or failure of the device is also more likely. Furthermore, the consideration of scaling effects is difficult in the development of control strategies for WECs (Wave Energy Scotland, 2016). In addition, the control of WECs should adapt to inevitable changes to the system dynamics over its life span, which are due to the ageing of components and marine growth. Similarly, the controller should maximize the availability of power generation by possibly adapting to non-critical subsystem failures and continuing operations until a suitable window is found for maintenance.

For these reasons, this thesis will address adaptive control strategies for WECs. Effective solutions in the area have been proposed by different research groups, for instance Fusco and Ringwood (2013), Bacelli (2014) and Zou *et al.* (2017). In this thesis, innovative approaches based on reinforcement learning and neural networks will be developed. This work will exploit the rapid improvements in the design of adaptive control schemes achieved by the robotics and computer science communities. Research in reinforcement learning was initiated by Pelamis Wave Power Ltd., but was stopped early on due to the bankruptcy of the company. This thesis presents the successful development of these strategies and an assessment of their performance.

1.2 Aims and objectives

1.2.1 Research question

Some types of control systems rely on a model of the plant to find the optimal control action. Plant models can be either explicit or implicit (Bordons and Camacho, 2007). Explicit models are created explicitly, often from physical processes and observations, whereas implicit models are built within the controller.

How to obtain an adaptive and optimal control performance that is independent of an explicit model of the plant?

1.2.2 Aim

As indicated by the research question that this thesis tries to address, the main aim of this project has been the development of adaptive optimal control strategies for the passive and active control of WECs, with a particular focus on model-free techniques. As a result, machine learning algorithms have been given special attention, exploiting the recent developments in the field.

1.2.3 Objectives

The over-arching aim of the project will be achieved via the following list of objectives:

1. First of all, an extensive literature review is carried out on the strategies developed to date for the control of WECs in order to analyse the properties of the state-of-the-art schemes and their possible short-comings.
2. The hydrodynamic model of a WEC is developed for the testing and validation of the proposed control algorithms. In particular, the model should present increasing complexity to assess the response under the influence of non-linear effects.

3. The performance of state-of-the-art control methods should be assessed using the generated WEC model.
4. The application of artificial neural networks to the active control of WECs should be investigated.
5. Suitable control algorithms should be developed for the passive and active control of WECs employing reinforcement learning. Care is required for the implementation of realistic force and displacement constraints.
6. Extensive experimental testing should be carried out to assess the performance of the developed algorithms against state-of-the-art control schemes. In particular, the ability of the controller to adapt to changes in sea state and system dynamics as well as its applicability to non-linear problems need to be analysed.

1.3 Contributions

1. The primary contribution of this thesis is the application of reinforcement learning to the time-averaged passive and active control of WECs. In particular, this work shows that:
   (a) Reinforcement learning learns the optimal parameters in every sea state for the maximization of power absorption, whilst abiding by displacement constraints and accounting for force saturation.
   (b) Learning time is acceptable for a realistic implementation.
   (c) The proposed control strategy is unaffected by system non-linearities and can adapt to changes in the system dynamics, e.g. due to marine growth or non-critical subsystem failure, since the approach does not rely on models to determine the control action.
   (d) The method has been tested with simulations of different WEC technologies considering a single unit. However, the modularity of its framework and its model-free nature enable the proposed strategy to be extended to arrays of WECs.
2. A further contribution is the application of reinforcement learning to the declutching control of a WEC. Monte-Carlo methods are used to determine the optimal timing for the application and release of the power take-off system force. The developed strategy is based on a simplistic implementation, but shows the potential for the development of real-time control schemes based on reinforcement learning for the control of WECs.
3. Finally, artificial neural networks are applied to the time-average active control of WECs. In particular, the learning time is found to be less than for reinforcement learning. Nevertheless, issues with constraints implementation need to be

addressed before practical use. Suggestions on how to exploit the advantages of this method and how to further improve its benefits are made.

1.3.1 Publications

1.3.1.1 Journal articles

- Anderlini, E., Forehand, D. I. M., Stansell, P., Xiao, Q., and Abusara, M. (2016). ”Control of a Point Absorber Using Reinforcement Learning”. *IEEE Transactions on Sustainable Energy*, 7, 4, October, pp. 1681-1690.
- Anderlini, E., Forehand, D. I. M., Bannon, E., and Abusara, M. (2017). ”Control of a Realistic Wave Energy Converter Model using Least-Squares Policy Iteration”. *IEEE Transactions on Sustainable Energy*, 8, 4, October, pp. 1618 - 1628.
- Anderlini, E., Forehand, D. I. M., Bannon E., and Abusara, M. (2017). ”Reactive Control of a Wave Energy Converter using Artificial Neural Networks.” *International Journal of Marine Energy*, 19, September, pp. 207-220.
- Anderlini, E., Forehand, D. I. M., Bannon, E., Xiao, Q. and Abusara, M. (2017). ”Reactive Control of a Two-Body Point Absorber using Reinforcement Learning”. *Ocean Engineering*, IDCORE special issue, in press.

1.3.1.2 Conference articles

- Anderlini, E., Forehand, D. I. M., Stansell, Bannon, E., Xiao, Q., and Abusara, M. (2016). ”Declutching Control of a Point Absorber based on Reinforcement Learning”. *Asian Wave and Tidal Energy Conference*, Singapore, October.
- Anderlini, E., Forehand, D. I. M., Bannon, E., and Abusara, M. (2017). ”Constraints Implementation in the Application of Reinforcement learning to the reactive control of a point absorber”. *Conference on Ocean, Offshore and Arctic Engineering*, Trondheim, June.
- Nambiar, A. Anderlini, E., Payne, G., Forehand, D. I. M., Kiprakis, A., and Wallace, R. (2017). ”Reinforcement Learning Based Maximum Power Point Tracking Control of Tidal Turbines”. *European Wave and Tidal Energy Conference*, Cork, August.

Note that in the last article some of the reinforcement learning strategies developed in this project have been applied to the control of a tidal turbine for the first time. The student has provided help with the development of the algorithms, with Dr Nambiar adapting them to the new application. Since this thesis deals with wave energy, the work done on tidal energy has not been included in this document, even though neural fitted Q-iteration is still presented in Chapter 4.

1.4 Organization of the thesis

The thesis comprises of seven additional chapters, which are outlined here.

Chapter 2 covers the development of dynamic models of wave energy converters. To start with, wave energy conversion is introduced and the functioning of WECs is described. Then, tools that are used for the modelling of WEC dynamics are discussed. Potential flow theory, in conjunction with the inclusion of some non-linear effects, is selected due to its associated good compromise in modelling quality and computational performance. The chapter continues with the description of the theory of wave-body interactions under the framework of potential flow theory. Equations of motions of the body are derived in both the frequency and time domains. Finally, the dynamics of three different point absorber geometries are modelled with systems of increasing complexity. These models will be used throughout this work for the assessment of control strategies.

Chapter 3 addresses the state-of-the-art methods for the control of wave energy converters. After a literature review, resistive and reactive control are analysed in detail. The model of a simple point absorber is employed as a case study to assess the performance of the control schemes.

In Chapter 4, reinforcement learning, a class of unsupervised learning algorithms, is described. After explaining its evolution from Markov decision processes, the issue of exploration and exploitation is discussed. Subsequently, Monte-Carlo methods, a type of reinforcement learning schemes, are introduced. This is followed by a detailed explanation of temporal difference schemes, which is an alternative, popular class of algorithms, including function approximation. In particular, Q-learning, Sarsa, neural fitted Q-iteration and least-squares policy iteration are described in detail. The chapter is concluded with the presentation of two classical examples of reinforcement learning applications.

In Chapter 5, an innovative strategy for the reactive control of WECs based on artificial neural networks is developed. After describing the proposed algorithm, a case study is presented for the assessment of the performance of the algorithm against state-of-the-art reactive control. The reader can found more information on artificial neural networks in Appendix A. The functioning of and the methods used to train neural networks are described here.

In Chapter 6, different reinforcement learning algorithms are used for the development of innovative control schemes for WECs. Firstly, Monte Carlo methods are applied to the declutching control of a WEC using a simple state-space description. The performance of the strategy is assessed with simulations of a point absorber, whose model was introduced in Chapter 2. Then, Q-learning, Sarsa and least-squares policy iteration are applied to the resistive and reactive control of WECs. Two case studies per control type are introduced to assess the performance of reinforcement learning against state-of-the-art control strategies.

Finally, in Chapter 7, the findings are summarised and conclusions are drawn. The document is completed by a discussion of proposed future work.

Chapter 2

Dynamic models of wave energy converters

WECs are machines that are designed to absorb part of the energy transported by water waves. The extracted power is not only dependent on the wave resource, but also on the physical properties and control of the device itself. An example of a WEC can be seen in Figure 2.1, which displays one of the devices developed by Pelamis Wave Power Ltd. This chapter describes methods that can be used to model the dynamics of WECs so that they can inform the design and control processes of these machines.

Firstly, the principle of wave energy conversion is summarised, and the main categories of WECs are presented. Then, different modelling tools for WEC dynamics are described. Of these, the theory of wave-body interactions based on potential flow is selected for the derivation of the equations of motion of the devices analysed in this thesis. This method is then described in detail and employed to simulate the dynamics of three machines, which will be studied throughout this work.

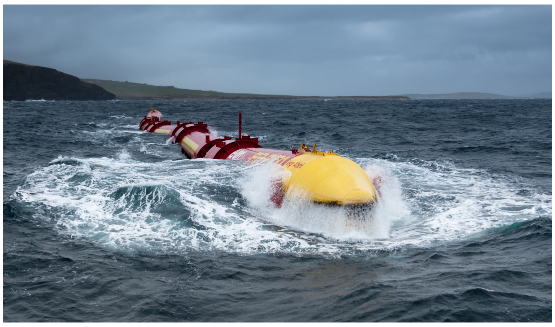


Figure 2.1: Pelamis P2 WEC in Orkney (with permission from Pelamis Wave Power Ltd.).

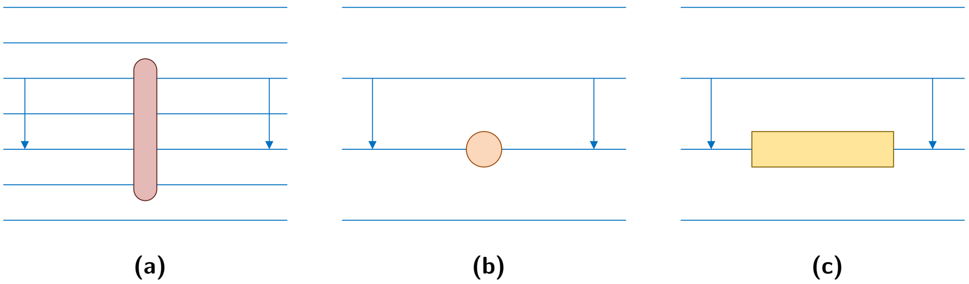


Figure 2.2: Three main categories of wave energy technologies: attenuator (a), point absorber (b) and terminator (c) in planar waves. Waves radiated by the WEC are not shown in this diagram for simplicity.

2.1 Wave energy conversion

Ever since WECs have been proposed by Salter (1974), numerous devices have been developed and built for the extraction of energy from gravity waves. A thorough description of wave energy designs and theory can be found in the books by Falnes (2005), Cruz (2008) and Korde and Ringwood (2016). Furthermore, Falnes (2007) presents a summary of the physics behind the functioning of WECs, while comprehensive reviews of the topic with a focus on the main technologies developed to date can be found in Drew *et al.* (2009), Falcão (2010), Titah-Benbouzid and Benbouzid (2014) and Santhosh *et al.* (2015). Moreover, a special issue of the Transactions of the Royal Society was dedicated to the field of wave energy (Farley *et al.*, 2012). In addition, Babarit *et al.* (2012) have compared the performance of the main WEC types using numerical studies. Finally, specialised reviews have been completed by Salter *et al.* (2002) and Falcão and Henriques (2016) on the PTO mechanisms and oscillating water columns, a WEC type, respectively.

Considering only WECs that extract kinetic energy from the ocean waves, i.e. excluding devices that rely on potential energy like hydro-power schemes such as Wave Dragon, all devices absorb power from the difference in motion between a *prime mover* that is excited by the incident waves and a *reference* (Korde and Ringwood, 2016). The reference may be represented by the sea floor, another body with a higher inertia and at greater depth or an internal moving mass. The prime mover need not be fixed, as it can consist in flexible membranes or even the water surface itself, as is the case for oscillating water columns. The treatment of all wave energy technologies developed to date goes beyond the scope of this thesis. Here, we first summarize the three main categories of WECs. Focusing on point absorbers, we then describe the different PTO mechanisms that have been developed for the extraction of energy from water waves.

Drew *et al.* (2009) classify WECs into three main categories: attenuators, point absorbers and terminators. Attenuators, shown in Figure 2.2a, are placed along the dominating direction of wave propagation, thus absorbing energy along their length. An example is the device produced by Pelamis Wave Power Ltd., which can be seen in Figure 2.1. Stansell and Pizer (2013) have demonstrated that attenuators present an energy absorption proportional to the displaced volume irrespective of volume constraints as opposed to point absorbers, whose performance is thus limited by their size. Therefore, they seem to be preferable from a perspective of economies of scale. Point absorbers are the second category that can be seen in Figure 2.2b. They present a characteristic length which is small with respect to the wave length. Due to their small size, wave direction is not important for these machine, which tend to have an axisymmetric geometry. Examples are the units produced by Ocean Power Technology or Carnegie Wave Power. Terminators represent the final WEC category and are shown in Figure 2.2c. They are placed perpendicular to the predominant wave direction and physically intercept the incoming waves. They can greatly benefit from economies of scale; however, terminators should be placed in near-shore locations, where waves are channelled along a main direction. The Salter Duck, first proposed in Salter (1974), and the Aquamarine Oyster are examples of this category.

In general, point absorbers present fewer joints than attenuators thus reducing the complexity of the problem. Additionally, although parametric coupling between pitch, roll and heave have been well documented (Falnes, 2005), the degrees of freedom of axisymmetric point absorbers (with axisymmetric PTO systems) can be decoupled if assuming linear wave theory. Therefore, they can be modelled with fewer degrees of freedom, which simplifies the understanding of the working principles and results in a smaller computational cost. For these reasons, only point absorbers have been analysed within this work, although the developed strategies should be applicable to non-linear systems. In particular, the focus has been on the study of individual devices. Nevertheless, the control strategies that are developed can be extended to the treatment of attenuators and terminators as well as arrays of WECs. In the next section, power take-off (PTO) systems are addressed in greater detail.

2.1.1 Power take-off systems

The main purpose of the PTO system is to transform the energy associated with the motions of the primary mover to a smooth flow of energy suitable for being delivered to the electrical grid. This poses significant challenges, since the oscillations of the device can be stochastic in irregular waves, whereas the electricity flowing in the national grid is sinusoidal and subject to quality checks (Cruz, 2008). Furthermore, wave energy varies on slower time scales as well, with more energy being available in storms and during the winter months. Nevertheless, it is the stringent requirements on the delivered

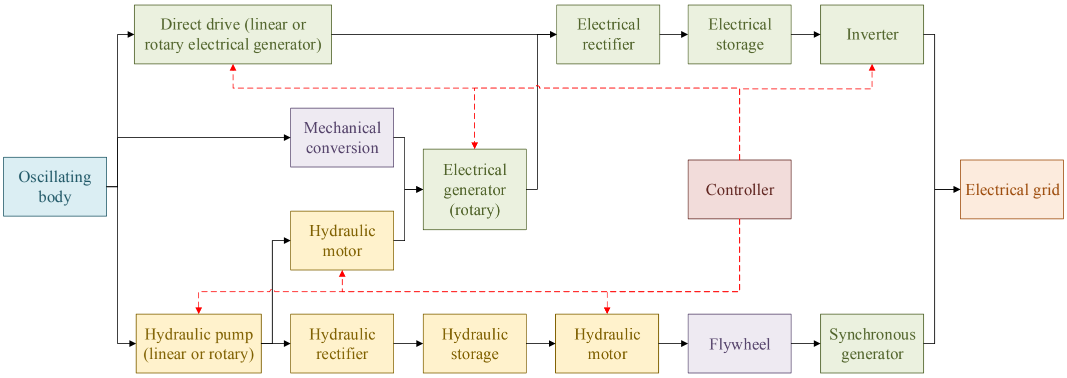


Figure 2.3: Block diagram of common PTO configurations, adapted from Bacelli (2014). Mechanical, hydraulic and electrical components are indicated by violet, yellow and green blocks, respectively.

electricity that are most relevant for the design of the PTO system. In addition, the direction of the flow is inverted every half wave cycle due to the oscillatory nature of gravity waves, which may create problems with existing generators. Therefore, since the first studies, storage systems and rectifiers have been included in the design, with hydraulic systems representing a clear candidate initially due to the technology transfer from oil and gas. Nowadays, it is possible to recognize four main types of PTO systems as displayed in Figure 2.3: hydraulic with hydraulic rectifier, hydraulic with electrical rectifier, mechanical with mechanical rectifier and direct drive.

In hydraulic PTO systems with a hydraulic rectifier, the hydraulic system provides good energy storage and flow rectification capacity (Forehand *et al.*, 2016). As a result, it is possible to use a simpler, cheaper synchronous generator. In a hydraulic PTO, the motion of the hydrodynamic absorbing body drives hydraulic fluid through rams in a hydraulic circuit with two parallel branches. A rectifying valve ensures the liquid flows in only one direction into the hydraulic motor, which is connected to a flywheel. High- and low-pressure accumulators result in a smooth hydraulic flow to the motor, while the flywheel, which is connected to the generator, further smooths out any oscillation. These systems have been implemented on most of the original WEC technologies, including by Pelamis Wave Power Ltd. (Henderson, 2006). Nowadays, a very efficient solution is proposed by Artemis Intelligent Power[1].

A similar unit is represented by the hydraulic PTO with electrical rectifier, where the hydraulic system is used only to increase the speed of the motions from the very low velocities associated with the motions of the device due to the wave excitation. Additionally, it enables the use of rotatory electrical generators, which are cheaper than

[1] http://www.artemisip.com/

linear ones. The energy storage function is provided by the electrical system through the use of batteries. In the electro-mechanical PTO, the increase in velocity is provided by a gearbox. These systems have the potential of lower capital and maintenance costs and increased efficiency due to the smaller number of components that may fail. An example is the PTO system designed by Umbra Cuscinetti (Castellini *et al.*, 2014).

Finally, direct-drive PTO systems have been proposed as well. These are machines that convert the kinetic energy of the moving body straight to electrical energy through a linear generator with permanent magnets. The design of the generator is very complex and performed ad-hoc in order to deal with the very low velocity of the wave motion ($<$ 10 m/s) and the associated large force (or torque for rotatory generators). The moving body is connected directly to the translator of the generator and power electronics are necessary to rectify the signal and provide power smoothing. An example of this PTO can be found in the Seabased point absorber, which has been developed at Uppsala University over a number of years (Danielsson, 2006; Eriksson, 2007; Waters, 2008; Stalberg *et al.*, 2008; Lejerskog *et al.*, 2015).

Some more exotic PTO designs have been proposed recently, e.g. through the use of piezo-electric membranes. An up-to-date description of the state-of-the-art technologies can be found in the website of Wave Energy Scotland dedicated to their PTO projects[2]. Additionally, Salter *et al.* (2002) and Peñalba Retes and Ringwood (2016) provide a review of the various PTO systems with a particular focus on their modelling. In this work, the PTO system is not modelled with the exception of the Seabased device, with an ideal PTO control force being employed instead for simplicity. Hence, the developed control strategies mainly deal with the control of the motions of the prime mover, which can be achieved in practice through the control of the PTO system using conventional strategies, such as PID control.

2.1.2 Background of dynamic modelling for wave energy converters

Although the control algorithms proposed as part of this work do not rely on explicit models of the system dynamics to select the control action, models of WECs have been developed in order to validate the schemes. The complexity of the simulation model also sets the severity of the control challenge. As a result, due to the real-time nature of the control strategies, models in the time domain are investigated. Although ideally non-linear hydrodynamic models should be employed to assess the performance of the learning algorithms, the duration of the learning process meant that the computational cost of the simulations would be excessive. Hence, linear models have been used instead, with the application of some non-linearities to the PTO models.

[2] `http://www.waveenergyscotland.co.uk/programmes/details/power-take-off/`

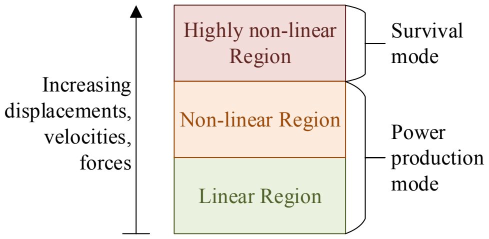


Figure 2.4: Operating regions for WECs and corresponding operational modes, adapted from Peñalba Retes *et al.* (2015).

Similarly, unfortunately, the project had no funding for experimental testing, which should be ideally used to validate the numerical results.

A review of numerical methods for the modelling of the dynamics of WECs can be found in Li and Yu (2012), with a particular focus on point absorbers. Peñalba Retes *et al.* (2015), Giorgi *et al.* (2016a) and Peñalba *et al.* (2017) treat the non-linear approaches more in detail, with the second work dealing mainly with point absorbers. As shown in Figure 2.4, it is possible to identify three main regions of operation for WECs: linear, non-linear and highly non-linear regimes. While the first two areas encompass the normal power generating mode, the last one is specific to the survival mode wave energy devices have to enter in extreme conditions. Over the years, tools have been developed for the modelling of WECs in the three regions, which present higher complexity and computational requirements the more non-linear effects are accounted for.

Linear methods have been investigated first due to their simpler nature and lower computational cost. These models are based on potential flow theory, carrying across knowledge from the field of marine hydrodynamics. Although many articles have been published on the topic, the books by Newman (1977) and Falnes (2005) represent the most reputable summaries of these methods. Whereas the former mainly deals with marine hydrodynamics from an offshore engineering and naval architecture perspective, the latter is specific to wave energy conversion. Thanks to the assumption of ideal, potential flow, linear methods allow the separation of the force components, so that it is possible to study their individual contribution to the dynamics of the system. Hence, the effect of the waves is divided into incident, diffraction and radiation components, as will be treated in the next section. These force components can be calculated analytically for specific geometries (Li and Yu, 2012), such as hemispheres and cylinders. However, for most realistic geometries, it is necessary to use boundary integral equation methods (Li and Yu, 2012), with commercial packages such as WAMIT (2013), Ansys Aqwa and open-source software NEMOH (Babarit and Delhommeau, 2015) being standard in the industry. These tools discretize the body surface with a number of panels in order to calculate the hydrodynamic coefficients. WAMIT can also remove the effect of irregular frequencies (Zhu and Lee, 1994), i.e. numerical errors associated with resonant frequencies dependent on the internal surface of the body. In addition, the hydrostatic and mooring restoring forces are modelled as linear functions of the body displacement and velocity. Initially, linear models were developed in the frequency-domain. Then, using Cummins (1962) formulation for the radiation force, the equations of motions of WECs were transformed to the time domain. The time-domain methods enable the development of superior control strategies for WECs (Peñalba Retes *et al.*, 2015), and they are fundamental for the modelling of non-linear PTO effects due to end stops (Pizer and Henderson, 2010).

The non-linear regime corresponding to the power absorbing mode can still be modelled with potential flow, although some modifications are necessary (Peñalba Retes *et al.*, 2015). The Froude-Krylov (i.e. incident wave) and hydrostatic forces are now treated as non-linear, calculating the actual forces at every time step based on the instantaneous wetted and water-plane areas, respectively. It is clear that this process can present an extremely high computational cost (Giorgi *et al.*, 2016a). Therefore, techniques have been proposed for the calculation of the coefficients using analytical formulae for established geometries (Giorgi *et al.*, 2016a). Similarly, at Pelamis Wave Power Ltd., the incident wave and hydrostatic coefficients were pre-calculated for a range of draughts and roll angles to speed up code performance (Pizer and Henderson, 2010). It is extremely important to notice, however, that non-linear effects on the hydrostatic force should be modelled only in conjunction with a non-linear Froude-Krylov force, since otherwise the code performance becomes worse than the linear code (Giorgi *et al.*, 2016a). This is because of the possible scenario of a WEC flying in air when no restoring force is applied. Furthermore, non-linear mooring effects can be included in the modelling of the WEC dynamics, using either a quasi-static or a fully dynamic approach (Harnois *et al.*, 2015). The latter results in a much higher computational cost, and it requires the discretization of the mooring lines in a number of finite elements. In addition, viscous damping effects, which can be significant for geometries with sharp edges, can be modelled using the Morison *et al.* (1950) equation. The drag coefficient used to be calculated from experimental measurements in wave basins, but can now be obtained from virtual wave basins using computational fluid dynamics (CFD) (Peñalba Retes *et al.*, 2015; Giorgi *et al.*, 2016a; Davidson *et al.*, 2016).

CFD represent the method of choice for the modelling of WECs under extreme wave conditions (Li and Yu, 2012; Peñalba Retes *et al.*, 2015). CFD approaches for WECs usually rely on the solution of the Navier-Stokes equations with two fluids using the finite-volume approach. Most commonly Reynolds-Averaged Navier-Stokes equations (RANS) solvers are employed, also relying on commercial software such Ansys CFX, Ansys Fluent, Star-CCM+ and (open-source) OpenFoam. In addition, smoothed-particle hydrodynamics (SPH) and other schemes more specific to wave energy can be employed as well. A description of these methods goes beyond the scope of this study, with these approaches treated in Li and Yu (2012) and Peñalba Retes *et al.* (2015).

In general, CFD approaches provide the best agreement with experimental measurements in tests with high, steep waves, with the quality of the prediction provided by the three approaches becoming similar in longer, gentler waves (Li and Yu, 2012; Peñalba Retes *et al.*, 2015). In particular, even the prediction provided by potential theory with some non-linear components was deemed of sufficient quality by Pelamis Wave Power Ltd. However, as aforementioned, the more complex the method, the longer the associated computational cost. In particular, Giorgi *et al.* (2016a) have shown that the non-linear methods present computational requirements an order of magnitude greater than those of linear methods, and CFD approaches up to 5 order of magnitude greater.

This project mainly deals with the development of control strategies for WECs, which are likely to require a relatively long time (12-24 hours) for convergence. For this reason, CFD approaches have been discarded. Furthermore, the PTO control force is expected to be non-linear. Therefore, we decided to employ mainly linear, time-domain methods to model the dynamics of WECs with a weakly non-linear PTO model. Although the proposed machine learning strategies rely on non-linear methods and their use is motivated by the possible improvement they can bring to the control of actual WEC devices, linear hydrodynamic models have been deemed to be a good platform for the initial development of the proposed strategies. Additionally, individual WECs are considered to further reduce the computational cost of the simulations, with the methods being extensible to the treatment of wave farms. Some non-linear effects, such as viscous drag and mooring forces, have been included in the analysis in order to assess the control behaviour under non-linear conditions. Although commercial software have been developed such as WaveDyn (Lucas *et al.*, 2012), InWave (Combourieu *et al.*, 2014) and (open-source) WEC-Sim (Ruehl *et al.*, 2014), which rely on potential flow theory and even include non-linear effects (Lawson *et al.*, 2014; Sirnivas *et al.*, 2016), we have preferred to develop the modelling tools from first principles in order to obtain a higher computational performance. In the next section, we describe the theory of wave-body interactions based on potential flow.

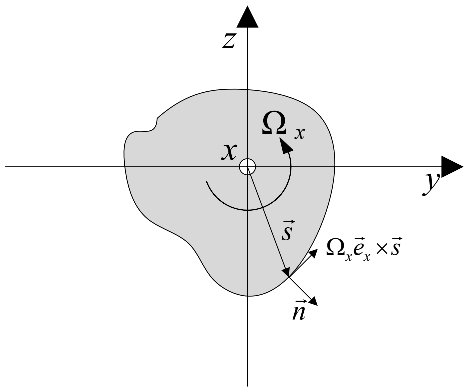


Figure 2.5: Rigid body with notation used in the derivation of its equations of motion.

Table 2.1: Modes of motion of a WEC.

Mode	Component	Direction	Name
1	x	along x-axis	surge
2	y	along y-axis	sway
3	z	along z-axis	heave
4	ϕ	about x-axis	roll
5	θ	about y-axis	pitch
6	ψ	about z-axis	yaw

2.2 Theory of wave-body interactions

The theory of wave-body interactions has its origin in the field of marine hydrodynamics. The same equations that will be used here to model WECs were initially developed to describe the motions of ships and offshore structures in waves. A full derivation of the equations presented hereafter can be found in Newman (1977) and, more specifically to WECs, Falnes (2005). Furthermore, the reader is referred back to the nomenclature of this thesis for the description of each symbol.

The WEC is modelled as a rigid body free to move in six degrees of freedom. These modes of motions are described in Table 2.1, while the coordinate system is shown in Figure 2.5.

The system of equations that describes a simple, uncontrolled WEC subject to linear motion is derived in the following sections in the frequency domain. The time-domain form is then obtained through Fourier transforms. As these sections are based on Newman (1977) and Falnes (2005), these references are no longer reported to aid readability.

2.2.1 Hydrodynamic model

The behaviour of a fluid is described by pressure and flow velocity. These properties can be calculated at any point in the fluid domain using the conservation principles for mass and momentum. In potential flow theory, an ideal fluid is assumed, i.e. *incompressible*, *irrotational* and *inviscid*.

For an incompressible fluid with constant density ρ, mass conservation states the equality of the rate of mass entering and exiting a system, and is expressed by the continuity equation

$$\nabla \cdot \vec{v} = 0, \tag{2.1}$$

where $\vec{v}(x, y, z, t)$ indicates the flow velocity vector. The conservation of momentum is represented by the Navier-Stokes equations

$$\rho \left(\frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v} \right) = \vec{f} - \nabla p + \mu \nabla^2 \vec{v}, \tag{2.2}$$

with $\vec{f} = [0, 0, -\rho g]^T$ being the gravitational force per unit volume, g the gravitational acceleration, p the fluid pressure and μ the fluid viscosity. CFD approaches find the fluid pressure and flow velocity by solving the system of equations represented by (2.1) and (2.2).

However, with the assumption of an inviscid, i.e. $\mu = 0$, and irrotational fluid, i.e. $\nabla \times \vec{v} = 0$, this is not necessary. Under these conditions, there exists a scalar function $\Phi(x, y, z, t)$ called *velocity potential* (hence, the name potential flow theory) such that

$$\vec{v} = \nabla \Phi. \tag{2.3}$$

By substituting (2.3) into the continuity equation (2.1), the Laplace equation is obtained

$$\nabla^2 \Phi = 0. \tag{2.4}$$

The velocity potential is obtained from the solution of (2.4), and then the flow velocity from (2.3).

The assumption of inviscid fluid is used to simplify (2.2) by dropping the last term. Therefore, it is possible to obtain the Bernoulli equation by integration (2.2) along a streamline of the velocity field:

$$\frac{p}{\rho} + \frac{\partial \Phi}{\partial t} + \frac{1}{2} (\nabla \Phi)^2 + gz = C, \tag{2.5}$$

with C being an integration constant.

Boundary conditions are necessary to calculate the velocity potential and pressure from

(2.4) and (2.5). If the rigid body is assumed to be *impermeable*, the velocity of a fluid particle on its surface must be zero in the direction normal to the body surface, where the direction is given by the unit normal vector $\vec{n}$ directed into the fluid domain. For a body moving with velocity $\vec{v}_{\text{b}}$, the kinematic boundary condition on the body surface is given by

$$\frac{\partial \Phi}{\partial \vec{n}} = \vec{v}_{\text{b}} \cdot \vec{n}. \tag{2.6}$$

For a static body, the boundary condition reduces to

$$\frac{\partial \Phi}{\partial \vec{n}} = 0. \tag{2.7}$$

Furthermore, assuming the sea floor to be planar and horizontal, the sea-floor boundary condition can be expressed as

$$\frac{\partial \Phi}{\partial z} = 0 \quad \text{on } z = -h, \tag{2.8}$$

where h is the water depth.

An additional kinematic condition must be specified on the free water surface, which is described by

$$z = \zeta(x, y, t), \tag{2.9}$$

where ζ is the wave elevation. The free-surface kinematic condition states that a fluid particle on the free-surface is assumed to remain on the free surface, and is then expressed as

$$\frac{\partial\left(z - \zeta(x, y, t)\right)}{\partial t} + \vec{v} \cdot \nabla\left(z - \zeta(x, y, t)\right) = 0 \quad \text{on } z = \zeta(x, y, t). \tag{2.10}$$

By substituting (2.3) into (2.10), an explicit version of the kinetic free surface boundary condition is obtained:

$$\frac{\partial \zeta}{\partial t} + \frac{\partial \Phi}{\partial x}\frac{\partial \zeta}{\partial x} + \frac{\partial \Phi}{\partial y}\frac{\partial \zeta}{\partial y} - \frac{\partial \Phi}{\partial z} = 0 \quad \text{on } z = \zeta(x, y, t). \tag{2.11}$$

On the free surface ($z = 0$), the water pressure equals the atmospheric pressure p_{atm}. Assuming the fluid to be motionless and neglecting the surface tension at the air-water interface, the constant of integration in the Bernoulli equation (2.5) is obtained as $C = p_{\text{atm}}/p$. Therefore, it is possible to define the dynamic free surface boundary condition as

$$g\zeta + \frac{\partial \zeta}{\partial t} + \frac{1}{2}\left[\left(\frac{\partial \Phi}{\partial x}\right)^2 + \left(\frac{\partial \Phi}{\partial y}\right)^2 + \left(\frac{\partial \Phi}{\partial z}\right)^2\right] = 0 \quad \text{on } z = \zeta(x, y, t). \tag{2.12}$$

2.2.1.1 Linear wave theory

On the one hand, the Laplace equation (2.4) expressing the velocity potential is linear; on the other hand, the free surface conditions (2.11) and (2.12) are non-linear. Linear wave theory is used to simplify these conditions so as to obtain a linear relationship between the motion of and forces acting on the rigid body and the wave amplitude. An intermediate step is to express the velocity potential as proportional to the wave elevation.

Equations (2.11) and (2.12) can be linearised by neglecting higher-order terms and by approximating the free-surface to $z = 0$ rather than (2.9). This assumes a small wave elevation as compared with the wave length. As a result, the linearised free-surface kinematic and dynamic boundary conditions become

$$\frac{\partial \zeta}{\partial t} = \frac{\partial \Phi}{\partial z} \quad \text{on } z = 0, \text{ and} \tag{2.13a}$$

$$g\zeta + \frac{\partial \Phi}{\partial t} = 0 \quad \text{on } z = 0, \tag{2.13b}$$

respectively. The combined free-surface boundary condition is thus

$$\frac{\partial^2 \Phi}{\partial t^2} + g\frac{\partial \Phi}{\partial z} = 0 \quad \text{on } z = 0. \tag{2.14}$$

2.2.1.1.1 Plane harmonic waves

For harmonically oscillating planar waves with angular frequency ω, the velocity potential can be expressed as

$$\Phi(x, y, z, t) = \operatorname{Re}\left(\hat{\Phi}(x, y, z)e^{i\omega t}\right). \tag{2.15}$$

By substituting the trial solution into (2.4), the velocity potential can be computed by solving the Laplace equation $\nabla^2\hat{\Phi} = 0$ using the method of the separation of variables. Falnes (2005) shows that for planar waves of infinite width, the following particular solution can be obtained using (2.8) and (2.14):

$$\hat{\Phi} = -\frac{g}{i\omega}\zeta_{\mathrm{a}}\frac{\cosh(kz + kh)}{\cosh(kh)}e^{-ikx\cos\beta - iky\sin\beta}, \tag{2.16}$$

where ζ_{a} is the amplitude of the wave elevation, $k = 2\pi/\lambda$ the wave number, with λ being the wave length and β the wave direction. The assumption of planar waves is realistic near the shoreline, particularly if the bathimetry and geography of an area ensures waves are channelled towards the coast (Holthuijsen, 2007; Cruz, 2008). Conversely, in deeper waters, wave spreading should be taken into account.

The wave number is related to the wave frequency by the dispersion relation

$$\omega^2 = gk\tanh(kh). \tag{2.17}$$

By reformulating the free-surface boundary condition (2.12) as

$$\hat{\zeta} = -\frac{i\omega}{g}\hat{\Phi}|_{z=0}, \tag{2.18}$$

the wave elevation can be calculated as

$$\hat{\zeta} = \zeta_\text{a}e^{-ikx}. \tag{2.19}$$

From (2.3) and (2.16), it is possible to calculate the fluid flow velocity vector as

$$\vec{v}_x = \frac{\partial\hat{\Phi}}{\partial x} = \omega\zeta_\text{a}\cos\beta\frac{\cosh(kz+kh)}{\sinh(kh)}e^{-ikxcos\beta-ikysin\beta}, \tag{2.20a}$$

$$\vec{v}_y = \frac{\partial\hat{\Phi}}{\partial x} = \omega\zeta_\text{a}\sin\beta\frac{\cosh(kz+kh)}{\sinh(kh)}e^{-ikxcos\beta-ikysin\beta}, \tag{2.20b}$$

$$\vec{v}_z = \frac{\partial\hat{\Phi}}{\partial x} = i\omega\zeta_\text{a}\frac{\sinh(kz+kh)}{\sinh(kh)}e^{-ikxcos(\beta)-ikysin\beta}. \tag{2.20c}$$

Using the linearised Bernoulli equation, i.e. ignoring $(\nabla\Phi)^2$ from (2.5), the pressure can be obtained as the sum of a hydrodynamic and hydrostatic term:

$$p - p_\text{atm} = -\underbrace{\rho\frac{\partial\Phi}{\partial t}}_{\text{dynamic}} - \underbrace{\rho g z}_{\text{static}}. \tag{2.21}$$

Therefore, it is possible to express the hydrodynamic pressure in complex form as a function of the velocity potential $\hat{\Phi}$ or the wave amplitude as

$$\hat{p} = -i\omega\rho\hat{\Phi} = \rho g\zeta_\text{a}\frac{\cosh(kz+kh)}{\cosh(kh)}e^{-ikxcos(\beta)-ikysin\beta}. \tag{2.22}$$

While (2.16) provides a linear relationship between the velocity potential and the wave elevation, the wave elevation and the hydrodynamic pressure are related by (2.22).

2.2.1.1.2 Dispersive waves

Gravity ocean waves are dispersive, i.e. their velocity of propagation is dependent on their frequency of oscillation. The velocity of propagation of a single wave is denoted by the phase velocity

$$v_\mathrm{p} = \frac{\omega}{k} = \frac{g}{\omega}\tanh(kh) \tag{2.23}$$

for constant depth h. Waves with different frequencies travel at different velocities. The group velocity describing the velocity of the envelope modulating of the dispersive wave is defined as

$$v_\mathrm{g} = \frac{\mathrm{d}\omega}{\mathrm{d}k} = \frac{g}{2\omega}\left(1 + \frac{2kh}{\sinh(2kh)}\right)\tanh(kh). \tag{2.24}$$

2.2.1.2 Forces acting on the rigid body

Defining the three-dimensional vectors of the hydrodynamic force and moment applied to the body as $\vec{f}$ and $\vec{m}$, respectively, the six-dimensional generalised force vector is expressed as

$$\boldsymbol{f} = \begin{bmatrix} \vec{f} \\ \vec{m} \end{bmatrix}. \tag{2.25}$$

The hydrodynamic forces and moments acting on the WEC are thus computed by integrating the pressure over the wetted surface area as

$$\boldsymbol{f} = -\iint_S p\boldsymbol{n}\mathrm{d}S, \tag{2.26}$$

where the six-dimensional normal vector is defined as

$$\boldsymbol{n} = \begin{bmatrix} \vec{n} \\ \vec{s}\times\vec{n} \end{bmatrix}. \tag{2.27}$$

The vector $\vec{s}$ expresses the position of the infinitesimal surface element $\mathrm{d}S$ with respect to the selected reference system, as shown in Figure . The velocity of the element is given by $\vec{v}_\mathrm{b} = \vec{u} + \vec{\Omega}\times\vec{s}$, with $\vec{u}$ and $\vec{\Omega}$ being the three-dimensional linear and angular velocity of the floating body, respectively. The six-dimensional generalised velocity vector can thus be expressed as

$$\boldsymbol{v} = \begin{bmatrix} \vec{u} \\ \vec{\Omega} \end{bmatrix}. \tag{2.28}$$

Assuming the motions of the rigid body to be small as compared with the wave amplitude, the velocity potential can be linearised and expressed as the sum of three contributions, namely due to wave *incidence* (Φ_i), *diffraction* (Φ_d) and *radiation* effects

(Φ_r):

$$\Phi = \underbrace{\Phi_\mathrm{i}}_{\text{incident}} + \underbrace{\Phi_\mathrm{d}}_{\text{diffracted}} + \underbrace{\Phi_\mathrm{r}}_{\text{radiated}} . \tag{2.29}$$

The first term describes the effect due to the unperturbed incident wave, while the second term expresses the disturbance to the incident wave due to the presence of the body. The radiation component describes the effects that the oscillating body exerts on a calm fluid (i.e. in the absence of incoming waves).

Substituting (2.15), (2.21) and (2.22) into (2.26) yields the complex generalised force vector

$$\hat{\boldsymbol{f}} = i\omega\rho \iint_S \hat{\Phi}\boldsymbol{n}\mathrm{d}S = \underbrace{i\omega\rho \iint_S \hat{\Phi}_\mathrm{i}\boldsymbol{n}\mathrm{d}S}_{\text{incident force}} + \underbrace{i\omega\rho \iint_S \hat{\Phi}_\mathrm{d}\boldsymbol{n}\mathrm{d}S}_{\text{diffraction force}} + \underbrace{i\omega\rho \iint_S \hat{\Phi}_\mathrm{r}\boldsymbol{n}\mathrm{d}S}_{\text{radiation force}} . \tag{2.30}$$

2.2.1.2.1 Radiation force

Owing to linearity, the velocity potential associated with the radiated waves $\hat{\Phi}_\mathrm{r}$ is given by the linear combination of the potentials associated with the waves radiated by oscillation in each mode. In addition, the radiation potential associated with each mode is proportional to the oscillation amplitude. Therefore, the radiation velocity potential can be expressed for the rigid body as

$$\hat{\Phi}_\mathrm{r} = \sum_{j=1}^{6} \phi_j \hat{v}_j = \boldsymbol{\phi} \cdot \hat{\boldsymbol{v}}, \tag{2.31}$$

where $\phi_j(x, y, z)$ is a function of position, but independent of time, and where $\hat{\boldsymbol{v}}$ is the complex generalised velocity vector.

The coefficients ϕ_j should be interpreted as the radiated velocity potential when the body oscillates in the j^th mode with unit velocity amplitude. Their calculation is the aim of the radiation problem, which is achieved by having $\hat{\Phi}_\mathrm{r}$ satisfy the Laplace equation (2.4) and the free-surface (2.14), sea-floor (2.8) and rigid-body (2.6) boundary conditions. In order to ensure a unique solution is found, an additional boundary condition at infinite distance from the body is specified based on the principle of conservation of energy:

$$\hat{\Phi}_\mathrm{r} \propto R^{-1/2} e^{-ikR}, \tag{2.32}$$

where $R \to \infty$ is the distance from the rigid body.

Substituting (2.31) into the radiation component of (2.30) yields the following l^th term of the radiation force vector $\hat{\boldsymbol{f}}$ when the body is forced to oscillate in the j^th mode only:

$$\hat{f}_{\mathrm{r},l} = i\omega\rho \iint_S \phi_j \hat{v}_j n_l \mathrm{d}S. \tag{2.33}$$

Since the body is assumed to be rigid, the terms of the generalized velocity vector $\hat{\boldsymbol{v}}$ are constant in the integration over the body surface. Hence, (2.33) can be re-written as

$$\hat{\boldsymbol{f}}_{\mathrm{r}} = -\boldsymbol{Z}(\omega)\hat{\boldsymbol{v}} \quad \text{in matrix form, or} \tag{2.34a}$$

$$\hat{f}_{\mathrm{r},l} = -Z_{l,j}\hat{v}_j, \qquad \text{with } Z_{l,j} = -i\omega\rho \iint_S \phi_j n_l \mathrm{d}S. \tag{2.34b}$$

This means that the radiation force is linearly proportional to the body velocity, with the constant of proportionality being given by $\boldsymbol{Z}$, known as the radiation impedance matrix. The negative sign indicates that the radiation force opposes the body motion. Applying the boundary condition on the body surface (2.6) results in the following expression for the radiation impedance:

$$Z_{l,j} = -i\omega\rho \iint_S \phi_j \frac{\partial \phi_l}{\partial n} \mathrm{d}S. \tag{2.35}$$

It is clear that the radiation matrix is complex and frequency-dependent. As a result, it is usually expressed as

$$\boldsymbol{Z}(\omega) = \boldsymbol{B}(\omega) + i\omega\boldsymbol{A}(\omega), \tag{2.36}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the frequency-dependent *added mass* and *hydrodynamic damping* matrices, respectively. When a body is accelerating in a fluid, some amount of the fluid must accelerate around the body as well. Therefore, greater force is required to accelerate the body in a fluid than in a vacuum. The added mass represents the extra inertia that needs to be added to the body in order to match the increase in force. In the case of a floating WEC, the added mass of the part of the body exposed to air is neglected. The hydrodynamic damping represents the damping effect water has on the oscillating body. However, this must not be confused with viscous damping, which is ignored by potential flow theory, since the fluid is assumed to be inviscid.

2.2.1.2.2 Excitation force

Similarly to the radiation problem, the incident and diffraction velocity potentials must satisfy the Laplace equation (2.4) and meet the boundary conditions on the free-surface (2.14), sea-floor (2.8), rigid-body (2.6) and at infinite distance from the body (2.32). The diffraction potential represents the disturbance caused by a force motion of the body with a resulting normal velocity equal and opposite to the velocity of the incident wave. Therefore, assuming that the incident velocity potential is linearly proportional to the wave amplitude, then so is the diffraction potential. Hence, the wave excitation force, which can be considered as the sum of the incident and diffraction wave force, can be expressed as a function of the wave elevation:

$$\hat{\boldsymbol{f}}_{\mathrm{e}} = \boldsymbol{H}(\omega, \beta)\hat{\zeta}, \tag{2.37}$$

where the vector of frequency- and direction-dependent excitation force coefficients is given by

$$\boldsymbol{H}(\omega, \beta) = i\omega\rho \iint_S \left(\hat{\Phi}_{\mathrm{i}}^0 + \hat{\Phi}_{\mathrm{d}}^0\right) \boldsymbol{n}\mathrm{d}S. \tag{2.38}$$

In (2.38), $\hat{\Phi}_{\mathrm{i}}^0$ and $\hat{\Phi}_{\mathrm{d}}^0$ are the incident and diffraction velocity potentials per wave unit amplitude, which can be obtained, for instance, by normalizing (2.16) with respect to the wave amplitude.

An alternative approach for the calculation of the excitation force was proposed by Haskind in 1957 (Newman, 1962):

$$\hat{f}_{\mathrm{e},j} = i\omega\rho \iint_S \left(\hat{\Phi}_{\mathrm{i}} \frac{\partial \phi_j}{\partial n} - \phi_j \frac{\partial \hat{\Phi}_{\mathrm{i}}}{\partial n}\right) \mathrm{d}S. \tag{2.39}$$

As it can be seen, this technique employs the solution of the radiation problem to find the excitation force. Not only does this method represent an effective tool for the validation of (2.37), but it is also more computationally efficient (WAMIT, 2013). Additionally, it can be run as part of the post-processing after the radiation coefficients have been calculated.

In irregular waves, the principle of superposition is employed to obtain the wave excitation force owing to the assumption of linearity. Thus, considering K individual waves each of which with elevation $\hat{\zeta}_k$, frequency ω_k and direction β_k, the total excitation force is given by

$$\hat{\boldsymbol{f}}_{\mathrm{e}} = \sum_{k=1}^{K} \boldsymbol{H}(\omega_k, \beta_k)\hat{\zeta}_k. \tag{2.40}$$

2.2.1.2.3 Hydrostatic force

The hydrostatic force represents the restoring force exerted by the water on a floating rigid body. By setting $p_{\mathrm{atm}} = 0$ in the Bernoulli equation (2.21), the hydrostatic component of the pressure can be calculated as

$$p = -\rho g z. \tag{2.41}$$

The generalized hydrostatic force vector is then computed by integrating the hydrostatic pressure over the body surface S:

$$\boldsymbol{f}_{\mathrm{h}} = \rho g \iint_S z\boldsymbol{n}\mathrm{d}S. \tag{2.42}$$

The relationship between the integral in (2.42) and the body position and attitude, i.e. orientation, with respect to an inertial reference frame is non-linear. Hence, linearisation will be necessary. Let us first define the six-dimensional vector ξ that describes the configuration of the body with respect to an inertial reference frame according to Table 2.1 as

$$\boldsymbol{\xi} = \begin{bmatrix} x & y & z & \phi & \theta & \psi \end{bmatrix}^T, \tag{2.43}$$

and the displacement from the equilibrium position $\boldsymbol{\xi}_0$ as $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \boldsymbol{\xi} - \boldsymbol{\xi}_0. \tag{2.44}$$

Assuming small perturbations, the hydrostatic force can be linearised with respect to the displacement vector $\boldsymbol{\eta}$ as follows:

$$\boldsymbol{f}_{\mathrm{h}} = -\boldsymbol{C}\boldsymbol{\eta}, \tag{2.45}$$

where $\boldsymbol{C} \geq \boldsymbol{0}$ is the *positive* semidefinite hydrostatic restoring coefficients matrix. The full derivation of the linearised form can be found in Newman (1977).

2.2.2 Multiple bodies

The equations presented above can be easily extended to the treatment of the interactions between N rigid bodies. Due to the assumption of linearization, the only change is in the number of components of the generalised vectors, which now becomes $6N$, i.e. 6 degrees of freedom per body. For instance, the term $A_{2,9}$ indicates the added mass the 1st body experiences in sway when the 2nd body is oscillating in heave. This notation follows the convention of WAMIT (2013).

Since restoring effects of one body should not interfere with other bodies, the overall restoring stiffness matrix is given by

$$\boldsymbol{C}_{\mathrm{h}} = \begin{bmatrix} \boldsymbol{C}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_2 & \dots & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{C}_n & \dots & \boldsymbol{0} \\ \vdots & \vdots & \dots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} & \dots & \boldsymbol{C}_N \end{bmatrix}, \tag{2.46}$$

where $\boldsymbol{C}_n$ is the 6-dimensional stiffness matrix of the n^{th} body.

2.2.3 Equation of motion in the frequency domain

The motion of a moving body is described by Newton's law as

$$\boldsymbol{M}\boldsymbol{a} = \sum \boldsymbol{f}, \tag{2.47}$$

where $\boldsymbol{M}$ is the mass matrix, $\boldsymbol{a}$ the six-dimensional generalised acceleration vector and $\sum \boldsymbol{f}$ indicates the sum of all forces and moments acting on the body. The mass matrix is dependent on the selected reference system. In the special case of system of reference being centred at the centre of gravity of the body, G, the mass matrix is given by

$$\boldsymbol{M} = \begin{bmatrix} m\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}^G \end{bmatrix}, \tag{2.48}$$

where m is the body mass and $\boldsymbol{I}^G$ indicates the inertia tensor with respect to G, while $\boldsymbol{I}$ is the three-dimensional identity matrix. Similarly to the stiffness matrix, the mass matrix for N bodies is given by

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_2 & \dots & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{M}_n & \dots & \boldsymbol{0} \\ \vdots & \vdots & \dots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} & \dots & \boldsymbol{M}_N \end{bmatrix}, \tag{2.49}$$

The forces acting on the body due to the interaction with the waves are due to wave excitation, radiation and hydrostatic restoring effects. Equation (2.47) may be extended to the treatment of N bodies, since the equations are linear in their variables. Considering the simplest case of a sinusoidal wave, (2.47) becomes

$$\boldsymbol{M}\boldsymbol{a} - \hat{\boldsymbol{f}}_{\text{r}} - \hat{\boldsymbol{f}}_{\text{h}} = \hat{\boldsymbol{f}}_{\text{e}}, \tag{2.50}$$

where $\boldsymbol{a} = i\omega\hat{\boldsymbol{v}}$. Substituting the results from (2.37), (2.34b) and (2.45) into (2.50) yields:

$$\boldsymbol{M}i\omega\hat{\boldsymbol{v}} + \boldsymbol{Z}(\omega)\hat{\boldsymbol{v}} + \boldsymbol{C}_{\text{h}}\frac{\hat{\boldsymbol{v}}}{i\omega} = \left[i\omega\left(\boldsymbol{M} + \boldsymbol{A}(\omega)\right) + \boldsymbol{B}(\omega) + \frac{\boldsymbol{C}_{\text{h}}}{i\omega}\right]\hat{\boldsymbol{v}} = \boldsymbol{H}(\omega, \beta)\hat{\zeta}. \tag{2.51}$$

Substituting $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{v}}/i\omega$ into (2.51) yields the following equation of motion in the frequency domain:

$$\{\left[\boldsymbol{C} - \omega^2 \left(\boldsymbol{M} + \boldsymbol{A}(\omega)\right)\right] + i\omega \boldsymbol{B}(\omega)\}\hat{\boldsymbol{\eta}} = \boldsymbol{H}(\omega, \beta)\hat{\zeta}. \tag{2.52}$$

In this work, a single wave direction is considered for simplicity: $\beta = 0$. Hence, effects due to wave spreading will no longer be treated in this thesis. From (2.52) it is possible to express the response amplitude operator (RAO), i.e. the response of the WEC(s) for unit wave amplitude, as

$$\frac{\hat{\boldsymbol{\eta}}}{\hat{\zeta}} = \{\left[\boldsymbol{C} - \omega^2 \left(\boldsymbol{M} + \boldsymbol{A}(\omega)\right)\right] + i\omega \boldsymbol{B}(\omega)\}^{-1}\boldsymbol{H}(\omega) \tag{2.53}$$

2.2.4 Equation of motion in the time domain

The equation of motion of a WEC in the time domain needs to be applicable to irregular waves as well as account for the memory effects in the fluid-body interaction due to radiated waves. Cummins (1962) was the first to propose a model that could describe the transient response thanks to the adoption of a test function for the radiation problem that comprises of two terms: while the first one considers the instantaneous effects of the fluid acceleration, the second one deals with the memory effects. The candidate solution to the radiation problem is thus

$$\Phi_{\mathrm{r}}(t) = \sum_{j=1}^{6N} \phi_j \ddot{\eta}_j(t) + \sum_{j=1}^{6N} \int_{-\infty}^{t} \chi_j(t-\tau)\dot{\eta}_j(\tau)\mathrm{d}\tau \tag{2.54}$$

subject to satisfying the Laplace equation (2.4) and the boundary conditions (2.8), (2.14), (2.32) and (2.12). While the first term represents the instantaneous hydrodynamic response, the convolution integral describes the memory effect, with χ being the impulse response. As a result of the velocity potential in (2.54), the radiation force according to Cummins (1962) is given by

$$\boldsymbol{f}_{\mathrm{r}}(t) = -\boldsymbol{\mu}\ddot{\boldsymbol{\eta}}(t) - \int_{-\infty}^{t} \boldsymbol{K}(t-\tau)\dot{\boldsymbol{\eta}}(\tau)\mathrm{d}\tau, \tag{2.55}$$

where $\boldsymbol{\mu} \in \Re^{6N\times 6N}$ is a matrix of constants is a matrix of constants and $\boldsymbol{K}(t) \in \Re^{6N\times 6N}$ is the radiation impulse response function matrix. $\boldsymbol{K}(t)$ is symmetric and its elements are zero for negative times, i.e. the relationship between the radiation force and the body velocity is causal (Falnes, 2005).

Similarly, the excitation force is also described by a convolution integral as described in Falnes (2005):

$$\boldsymbol{f}_{\mathrm{e}}(t) = \int_{-\infty}^{\infty} \boldsymbol{h}(t-\tau)\zeta(\tau)\mathrm{d}\tau. \tag{2.56}$$

However, as discussed in Falnes (1995), the excitation force is non-causal as opposed to the radiation force, i.e. $\boldsymbol{f}_{\mathrm{e}}(t_0)$ also depends on future values of the wave elevation $\zeta(t)$, for $t > t_0$. Furthermore, the excitation impulse response vector $\boldsymbol{h}(t) \in^{6N \times N}$ is different from zero for negative times.

In fact, the excitation force coefficients in the frequency domain, $\boldsymbol{H}(\omega)$, is the Fourier transform of the excitation impulse response function $\boldsymbol{h}(t)$. Similarly, by applying the Fourier transform to (2.55), it is possible to obtain the Ogilvie (1964) relations

$$\boldsymbol{B}(\omega) = \int_0^\infty \boldsymbol{K}(t) \cos \omega t \mathrm{d}t, \tag{2.57a}$$

$$\boldsymbol{A}(\omega) = \boldsymbol{\mu} - \frac{1}{\omega} \int_0^\infty \boldsymbol{K}(t) \sin \omega t \mathrm{d}t. \tag{2.57b}$$

As the impulse response $\boldsymbol{K}(t)$ is square integrable, the following equations hold true (Kristiansen *et al.*, 2006):

$$\lim_{\omega \to \infty} \int_0^\infty \boldsymbol{K}(t) \cos \omega t \mathrm{d}t = \lim_{\omega \to \infty} \int_0^\infty \boldsymbol{K}(t) \sin \omega t \mathrm{d}t = \mathbf{0}, \tag{2.58a}$$

$$\lim_{\omega \to \infty} \boldsymbol{B}(\omega) = \mathbf{0}, \tag{2.58b}$$

$$\lim_{\omega \to \infty} \boldsymbol{A}(\omega) = \boldsymbol{A}(\omega) = \boldsymbol{\mu}. \tag{2.58c}$$

Hence, the matrix of constants $\boldsymbol{\mu}$ in (2.55) is equal to the added mass at infinite wave frequency.

Similarly, using inverse Fourier transforms it is possible to show that (Falnes, 2005)

$$\boldsymbol{K}(t) = \frac{2}{\pi} \int_0^\infty \boldsymbol{B}(\omega) \cos \omega t \mathrm{d}t, \tag{2.59a}$$

$$\boldsymbol{K}(t) = -\frac{2}{\pi} \int_0^\infty \omega \left(\boldsymbol{A}(\omega) - \boldsymbol{A}(\infty)\right) \sin \omega t \mathrm{d}t. \tag{2.59b}$$

The integral in (2.59a) converges faster than the one in (2.59b); hence, (2.59a) is typically employed to calculate the radiation impulse response function using frequency-domain coefficients.

In conclusion, the time-domain equation of motion corresponding to (2.52) is given by

$$\left(\boldsymbol{M} + \boldsymbol{A}(\infty)\right) \ddot{\boldsymbol{\eta}}(t) + \int_\infty^t \boldsymbol{K}(t-\tau) \dot{\boldsymbol{\eta}}(\tau) \mathrm{d}\tau + \boldsymbol{C}\boldsymbol{\eta}(t) = \int_{-\infty}^\infty \boldsymbol{h}(t-\tau) \zeta(\tau) \mathrm{d}\tau. \tag{2.60}$$

2.2.5 Wave elevation generation

From (2.56), it is clear that the wave elevation time series is required in order to obtain the wave excitation force. For simplicity, here we consider *planar*, *linear* waves with no spreading. Additionally, the body is assumed to be sited at the centre of the reference system, about which point all forces and moments are taken.

2.2.5.1 Regular waves

The time-dependent wave elevation of a planar, sinusoidal wave with amplitude ζ_a and period T is given by (Holthuijsen, 2007)

$$\zeta(t) = \zeta_\mathrm{a} sin(\omega t + kx + \gamma), \tag{2.61}$$

where the circular wave frequency is $\omega = 2\pi/T$ and γ in this case represents a phase. Assuming $\gamma = 0$ and remembering that $x = 0$ at the origin, (2.61) can be expressed as

$$\zeta(t) = \zeta_\mathrm{a} R(t)\sin(\omega t), \tag{2.62}$$

where R is a ramp function that prevents divergence of the numerical solution of teh system dynamics during the initial transient response. This function is taken from NREL (2015) and is expressed as

$$R(t) = \begin{cases} 0.5\left[1+\cos\left(\pi + \dfrac{\pi t}{t_\mathrm{r}}\right)\right] & \text{if } t < t_\mathrm{r} \quad (2.63\text{a}) \\ 1 & \text{otherwise,} \quad (2.63\text{b}) \end{cases}$$

where t_r is the specified duration of the initial ramp function.

2.2.5.2 Irregular waves

In linear wave theory, the wave elevation in irregular waves can be obtained as the superposition of n_w individual sinusoidal waves. The content in real random waves is typically represented through a wave spectrum $S(\omega)$ (Holthuijsen, 2007), from which the individual wave components can be determined. Here only two of the most common wave spectrum models are employed: Bretschneider and JONSWAP. While the former is more representative of oceanic waters, e.g. off the West coast of Scotland, the latter, which stands for JOint North Sea WAve Project, is more suitable for shallower, enclosed seas, such as the North Sea (Holthuijsen, 2007).

The wave spectra equations are taken from Det Norske Veritas (2010), where the Bretschneider spectrum is labelled as Pierson-Moskowitz. The Bretschneider and JONS-

WAP spectra are obtained as

$$S_{\mathrm{B}}(\omega) = \frac{5}{16} H_{\mathrm{s}}^{2} \frac{\omega_{\mathrm{p}}^{4}}{\omega^{5}} \exp\left[-\frac{5}{4}\left(\frac{\omega}{\omega_{\mathrm{p}}}\right)^{-4}\right], \tag{2.64a}$$

$$S_{\mathrm{J}}(\omega) = (1 - 0.287 \ln \gamma) \frac{5}{16} H_{\mathrm{s}}^{2} \frac{\omega_{\mathrm{p}}^{4}}{\omega^{5}} \exp\left[-\frac{5}{4}\left(\frac{\omega}{\omega_{\mathrm{p}}}\right)^{-4}\right] \gamma^{\exp\left[-0.5 \frac{\omega - \omega_{\mathrm{p}}}{\sigma(\omega)\omega_{\mathrm{p}}}\right]}, \tag{2.64b}$$

respectively, where H_{s} is the significant wave height, $\omega_{\mathrm{p}} = 2\pi/T_{\mathrm{p}}$ is the peak circular wave frequency and T_{p} the peak wave period, $\gamma = 3.3$ is a non-dimensional peak shape parameter (Det Norske Veritas, 2010) and

$$\sigma(\omega) = \begin{cases} 0.07 & \text{if } \omega \leq \omega_{\mathrm{p}} \ , & (2.65a) \\ 0.09 & \text{otherwise.} & (2.65b) \end{cases}$$

The JONSWAP is a peakier spectrum and it is expected to be valid for the range $3.6 < T_{\mathrm{p}}/H_{\mathrm{s}} < 5$ (with T_{p} and H_{s} expressed in s and m, respectively) (Det Norske Veritas, 2010).

Once the wave spectrum is known for n_{w} individual values of the circular wave frequency, with a circular wave frequency step of $\Delta\omega$, the amplitude of each wave component is determined as (Holthuijsen, 2007)

$$a(\omega) = \sqrt{2S(\omega)\Delta\omega}. \tag{2.66}$$

Furthermore, each individual wave presents a different, random phase γ_i. Hence, the wave elevation in irregular waves is obtained as

$$\zeta(t) = R(t) \sum_{i=1}^{n_{\mathrm{w}}} a(\omega_i) \sin(\omega_i t + \gamma_i), \tag{2.67}$$

2.2.5.3 Obtaining sea state parameters

Energy content in waves usually varies every 0.5 to 6 hours, with average wave conditions being labelled as sea states (Holthuijsen, 2007). The wave height and period are two parameters that are used to define average wave conditions. As aforementioned, the significant wave height is the statistical quantity used to characterise the wave height. Its value corresponds to the mean of the 33% observed highest wave heights per sea state, since this corresponds approximately to the wave height that could be recorded by an observer on a ship (Holthuijsen, 2007). Furthermore, the significant wave height can also be computed more formally from the 0^{th} moment of the wave spectrum

(Holthuijsen, 2007):

$$H_\mathrm{s} = 4\sqrt{\int_0^\infty S(\omega)\mathrm{d}\omega}. \tag{2.68}$$

The peak wave period is the period corresponding to the maximum point of the wave period spectrum. Like the significant wave height, it can be calculated formally as (Holthuijsen, 2007)

$$T_\mathrm{p} = \frac{\left(\int_0^\infty \omega^{-2} S(\omega)\mathrm{d}\omega\right)\left(\int_0^\infty \omega S(\omega)\mathrm{d}\omega\right)}{\left(\int_0^\infty S(\omega)\mathrm{d}\omega\right)^2}. \tag{2.69}$$

Although the peak wave period is fundamental for the generation of waves using wave spectra models, the energy and mean zero-crossing wave periods are more commonly used when extracting information about the current sea state using wave buoys (Holthuijsen, 2007). While the former represents the period of regular waves having the same energy content as the analysed sea state, the latter indicates the mean period of waves crossing $\zeta = 0$. Similarly to H_s and T_p, they can be calculated as

$$T_\mathrm{e} = \frac{\int_0^\infty \omega^{-1} S(\omega)\mathrm{d}\omega}{\int_0^\infty S(\omega)\mathrm{d}\omega}, \tag{2.70a}$$

$$T_\mathrm{z} = \sqrt{\frac{\int_0^\infty S(\omega)\mathrm{d}\omega}{\int_0^\infty \omega^2 S(\omega)\mathrm{d}\omega}}, \tag{2.70b}$$

respectively (Holthuijsen, 2007). In particular, the energy wave period is usually most accurate, since it is least affected numerical errors associated with very small wave frequencies (Holthuijsen, 2007). Realistic sea states have a significant wave height ranging from 0 to 12 m and an energy wave period ranging from 5 to 18 s (Holthuijsen, 2007), although there may be exceptions.

The extraction of a wave spectrum from a time-domain wave trace is achieved through spectral analysis and the use of Fourier transforms as described in Appendix C of Holthuijsen (2007).

2.2.6 Approximation of the radiation force

Considering a practical implementation, the solution of the radiation and excitation convolution integrals can be very computationally demanding. On the one hand, the excitation force may be pre-generated for all time steps of the analysed wave trace, since it depends on the wave elevation signal. On the other hand, this is not possible for the radiation force, since it is dependent on the body velocity vector. Therefore, for the modelling of WECs, it is standard practice to approximate the radiation convolution integral to speed up the simulations (Ringwood *et al.*, 2014; Korde and Ringwood, 2016). In particular, the benefit of approximating the radiation convolution with a state-space approach is discussed in Taghipour *et al.* (2008) in terms of computational resources. This improvement is due to the Markovian property of state-space models, where the current state of the model summarises all past information (Pérez and Fossen, 2008). Pérez and Fossen (2008) present different frequency- and time-domain methods for the approximation of the radiation force based on system identification. Here, frequency-domain system identification is used, as described in Forehand *et al.* (2016).

Using inverse Fourier transforms, the frequency-domain radiation impedance function may be obtained as

$$\boldsymbol{K}(\omega) = \int_0^\infty \boldsymbol{K}(t) e^{-i\omega t} \mathrm{d}\omega = \boldsymbol{B}(\omega) + i\omega \left[\boldsymbol{A}(\omega) - \boldsymbol{A}(\infty)\right]. \tag{2.71}$$

Matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{K}$ are all symmetric (Falnes, 2005). Each radiation impedance function $K_{i,j}(\omega)$ is then fitted with a rational transfer function $\hat{K}_{i,j}(\omega)$ with polynomials in the numerator and denominator of order m and n, respectively (Pérez and Fossen, 2008; Forehand *et al.*, 2016):

$$\hat{K}_{i,j}(s, \boldsymbol{\theta}) = \frac{P(s,\boldsymbol{\theta})}{Q(s,\boldsymbol{\theta})} = \frac{p_m s^m + p_{m-1} s^{m-1} + \cdots + p_0}{s^n + q_{n-1} s^{n-1} + \cdots + q_0} \tag{2.72}$$

where $m < n$, $s = i\omega$ as per standard control literature (Franklin *et al.*, 2008) (not to be confused with states in Chapter 4) and the vector of parameters is defined as

$$\boldsymbol{\theta} = \begin{bmatrix} p_m & \dots & p_1 & q_{n-1} & \dots & q_0 \end{bmatrix}^T . \tag{2.73}$$

A first approximation to the parameters vector is found using least-squares error fitting (Levy, 1959). Then, this result is employed as the initial starting point for a second algorithm that relies on a damped Gauss-Newton iterative search method (Dennis and Schnabel, 1983). This second step is required to guarantee the stability of the computed transfer function (Forehand *et al.*, 2016), i.e. it ensures that its poles, i.e. the zeros of the denominator polynomial, are all in the left half-plane (Franklin *et al.*, 2008). From a practical perspective, (2.72) is implemented in Matlab using the *invfreqs* function. The orders m and n are increased incrementally for each radiation impedance function until the root-mean-square error between $K_{i,j}$ and $\hat{K}_{i,j}$ is less than 1%.

Once all approximate transfer functions have been obtained for all degrees of freedom (hence, $6N \times 6N$ in total, although symmetry greatly reduces the number of individually distinct transfer functions), the system is converted into a *single* equivalent state-space model. However, for each set of radiation convolution terms there is an infinite set of equivalent state-space models, most of which will be numerically unstable (Forehand *et al.*, 2016). Even formulations that have been constructed to be mathematically stable may become unstable as a result of number overflow and truncation, which can result in a change in the roots of the characteristic equation (eigenvalues of the system matrix). In the application to the approximation of the convolution integral, Forehand *et al.* (2016) have found a modified version of the controllable canonical form to suffer from this problem. For this reason, they have proposed the more robust approach of first converting each transfer function to zero-pole-gain form by factorizing the numerator and denominator polynomials. Subsequently, the resulting system of zero-pole-gain models is converted to a single state-space system in modal canonical form (Franklin *et al.*, 2008). Hence, the resulting approximate radiation force is given by

$$\dot{\boldsymbol{x}}_{\mathrm{ss}}(t) = \boldsymbol{A}_{\mathrm{ss}}\boldsymbol{x}_{\mathrm{ss}}(t) + \boldsymbol{B}_{\mathrm{ss}}\boldsymbol{u}_{\mathrm{ss}}(t), \tag{2.74a}$$

$$\int_{-\infty}^{t} \boldsymbol{K}(t-\tau)\dot{\boldsymbol{\eta}}(\tau)\mathrm{d}\tau \approx \boldsymbol{y}_{\mathrm{ss}}(t) = \boldsymbol{C}_{\mathrm{ss}}\boldsymbol{x}_{\mathrm{ss}}(t) + \boldsymbol{D}_{\mathrm{ss}}\boldsymbol{u}_{\mathrm{ss}}(t), \tag{2.74b}$$

where $\boldsymbol{u}_{\mathrm{ss}} = \dot{\boldsymbol{\eta}}$, $\boldsymbol{x}_{\mathrm{ss}}$ and $\boldsymbol{y}_{\mathrm{ss}}$ are the input, state and output vectors, and $\boldsymbol{A}_{\mathrm{ss}}$, $\boldsymbol{B}_{\mathrm{ss}}$, $\boldsymbol{C}_{\mathrm{ss}}$ and $\boldsymbol{D}_{\mathrm{ss}}$ are the state, input, output, and feedthrough matrices, respectively. In the case of the analysed point absorber geometries, $\boldsymbol{D}_{\mathrm{ss}} = \boldsymbol{0}$.

2.2.7 Additional non-linear forces

The time-domain model in (2.60) is based on linear hydrodynamics and ignores some realistic non-linear effects. In particular, the following effects can have a significant effect on the WEC dynamics and may be added to the previously obtained model based on potential flow theory:

- PTO system force: this is the force exerted by the PTO system onto the WEC, whose generalised vector is represented by $\boldsymbol{f}_{\mathrm{PTO}}$. This force represents the main control input. As such, it will be treated in the next chapter.
- Mooring force: this is the restoring force exerted by the mooring system onto the WEC, which will be treated in Section 2.2.7.1.
- Viscous damping force: this is a damping force caused by hydrodynamic viscous effects. Although the fluid is considered inviscid by potential wave theory, water viscosity can be significant in highly energetic waves and for particular device geometries (Cruz, 2008). A simple, but effective model for the treatment of viscous effects in conjunction with potential flow theory was proposed by Morison *et al.* (1950) and is treated in Section 2.2.7.2.
- Non-linear hydrostatic force: although a linearised hydrostatic force is employed in (2.60), in fact the restoring effects are dependent on the instantaneous waterplane area.
- Non-linear Froude-Krylov force: like the hydrostatic force, the actual Froude-

Krylov, or wave excitation force, changes based on the instantaneous underwater geometry.

Parametric approaches have been proposed for the treatment of non-linear hydrostatic and Froude-Krylov forces in a computationally efficient manner, as for instance discussed by Giorgi *et al.* (2016a). A similar approach was adopted at Pelamis Wave Power Ltd. for the development of a more realistic model of the WEC dynamics in high waves in conjunction with strip theory (Pizer and Henderson, 2010). Although the adoption of these techniques results in a prediction of the WEC motion of higher quality, only linear Froude-Krylov and hydrostatic forces have been considered to lower the computational cost of the simulations.

2.2.7.1 Mooring forces

Mooring forces can be analysed with different techniques of increasing complexity and quality of prediction: linear model, quasi-static and dynamic (Harnois *et al.*, 2015). Although non-linear mooring forces have been considered for control applications, for instance by Richter *et al.* (2013), here only a simplified, linear mooring force is analysed (Bacelli, 2014; Korde and Ringwood, 2016):

$$\boldsymbol{f}_{\mathrm{m}} = -\boldsymbol{B}_{\mathrm{m}}\dot{\boldsymbol{\eta}} - \boldsymbol{C}_{\mathrm{m}}\boldsymbol{\eta}, \tag{2.75}$$

where $\boldsymbol{f}_{\mathrm{m}}$ is the generalised mooring force vector and $\boldsymbol{B}_{\mathrm{m}}$ and $\boldsymbol{C}_{\mathrm{m}}$ the mooring damping and stiffness matrices, respectively.

2.2.7.2 Viscous damping force

Viscous damping effects can be modelled with the simple model proposed by Morison *et al.* (1950):

$$\boldsymbol{f}_{\mathrm{d}} = -\frac{1}{2}\rho\boldsymbol{C}_{\mathrm{d}}A_{\mathrm{d}}|\dot{\boldsymbol{\eta}} - \boldsymbol{v}_{\mathrm{G}}|\left(\dot{\boldsymbol{\eta}} - \boldsymbol{v}_{\mathrm{G}}\right), \tag{2.76}$$

where $\boldsymbol{C}_{\mathrm{d}}$ is the diagonal matrix of the drag coefficients, A_{d} is a characteristic area specific to each body and $\boldsymbol{v}_{\mathrm{G}}$ is the generalized vector of the unperturbed flow velocity at the centre of gravity of each body. Experimental tests should be performed to estimate the drag coefficients, as for instance done by Lok *et al.* (2014). However, identifying $\boldsymbol{C}_{\mathrm{d}}$ in a numerical wave tank with CFD as in Bhinder *et al.* (2011) may speed up and reduce the cost of the process (Giorgi *et al.*, 2016a).

2.2.8 Time-domain, dynamic model of wave energy converters

Using (2.74a), (2.74b), (2.75) and (2.76), the equation of motion in time domain in (2.60) may be re-written as

$$(\boldsymbol{M} + \boldsymbol{A}(\infty))\,\ddot{\boldsymbol{\eta}} + \boldsymbol{B}_{\mathrm{m}}\dot{\boldsymbol{\eta}} + \boldsymbol{C}_{\mathrm{ss}}\dot{x}_{\mathrm{ss}} + (\boldsymbol{C} + \boldsymbol{C}_{\mathrm{m}})\,\boldsymbol{\eta} = \boldsymbol{f}_{\mathrm{e}} + \boldsymbol{f}_{\mathrm{PTO}} + \boldsymbol{f}_{\mathrm{d}}, \tag{2.77a}$$

$$\dot{\boldsymbol{x}}_{\mathrm{ss}} = \boldsymbol{A}_{\mathrm{ss}}\boldsymbol{x}_{\mathrm{ss}} + \boldsymbol{B}_{\mathrm{ss}}\dot{\boldsymbol{\eta}}, \tag{2.77b}$$

where the time dependence of the generalized vectors has been dropped to simplify the notation. The reader is reminded that in (2.77a) and (2.77b) $\boldsymbol{M}$ indicates the inertia matrix of the WEC(s), $\boldsymbol{A}(infty)$ the added mass matrix at infinite wave frequency, $\boldsymbol{B}_{\mathrm{m}}$ and $\boldsymbol{C}_{\mathrm{m}}$ the damping and stiffness matrices associated with the mooring, respectively, $\boldsymbol{C}$ the hydrostatic restoring stiffness matrix, $\boldsymbol{f}_{\mathrm{e}}$ the excitation force vector, $\boldsymbol{f}_{\mathrm{PTO}}$ the PTO force vector and $\boldsymbol{f}_{\mathrm{d}}$ the viscous damping force vector. Additionally, $\boldsymbol{\eta}$ is the vector of the displacement of the WEC(s) in all degrees of freedom and $\boldsymbol{x}_{\mathrm{ss}}$ the vector of the state-space system that is used to approximate the convolution integral associated with the radiation force.

Throughout this work, the system of the equations of motion has been solved numerically. This requires the addition the additional variable $\boldsymbol{\nu}(t) = \dot{\boldsymbol{\eta}}(t)$. Rearranging (2.77a) and (2.77b), it is possible to obtain

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{P}\boldsymbol{x}(t) + \boldsymbol{Q}\boldsymbol{u}(t), \text{ where} \tag{2.78a}$$

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{\eta}^T & \boldsymbol{\nu}^T & \boldsymbol{x}_{\mathrm{ss}}^T \end{bmatrix}^T, \tag{2.78b}$$

$$\boldsymbol{u} = \boldsymbol{f}_{\mathrm{e}} + \boldsymbol{f}_{\mathrm{PTO}} + \boldsymbol{f}_{\mathrm{d}}, \tag{2.78c}$$

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ -(\boldsymbol{M} + \boldsymbol{A}(\infty))^{-1}(\boldsymbol{C} + \boldsymbol{C}_{\mathrm{m}}) & -(\boldsymbol{M} + \boldsymbol{A}(\infty))^{-1}\boldsymbol{B}_{\mathrm{m}} & -(\boldsymbol{M} + \boldsymbol{A}(\infty))^{-1}\boldsymbol{C}_{\mathrm{ss}} \\ \boldsymbol{0} & \boldsymbol{B}_{\mathrm{ss}} & \boldsymbol{A}_{\mathrm{ss}} \end{bmatrix}, \tag{2.78d}$$

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{0} \\ (\boldsymbol{M} + \boldsymbol{A}(\infty))^{-1} \\ \boldsymbol{0} \end{bmatrix}. \tag{2.78e}$$

Equation (2.78) can then be solved numerically using any from a range of well documented numerical solvers (Süli and Mayers, 2003). In this work, fixed-step solvers have been adopted, since the developed control schemes could have been implemented on a real system if there had been sufficient time and resources for experimental testing. In particular, either a first-order-accurate Euler scheme or a fourth-order-accurate Runge-Kutta method have been employed. The former was found to result in sufficient accuracy with the linear models, whereas the latter was preferred with the weakly non-linear model of the PTO unit.

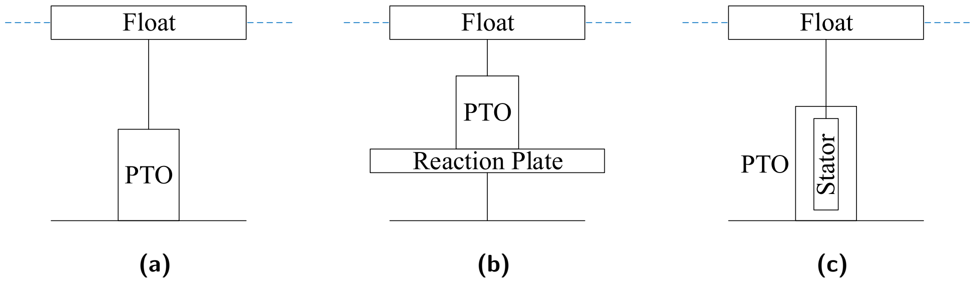


Figure 2.6: Diagram of the point absorbers simulated in this thesis.

2.3 Modelled wave energy converters

As aforementioned, in this thesis only point absorbers are considered. In particular, three distinct devices have been analysed as shown in Figure 2.6: a simple float that reacts against the sea floor (Figure 2.6a), a two-body point absorber that comprises of a float and a reaction plate (Figure 2.6b), and a float that reacts against a moving mass within a PTO system sited at the sea bottom (Figure 2.6c). The difference between the first and last devices consists in the model of the PTO system, which is represented by a realistic non-linear model in the third study. The last two devices have been inspired from actual WECs, while the former presents a simple case study first introduced by Newman (1977).

In this chapter, the motions of the three machines are modelled as described above. The commercial software WAMIT (2013) is used to extract the respective hydrodynamic coefficients assuming deep water. Furthermore, the response of each device in both regular and irregular waves is shown when no control force is applied. The wave elevation of the two traces can be seen in Figure 2.11 and Figure 2.12 for regular waves with unit amplitude and a period of 8 s and irregular waves with a Bretschneider spectrum and H_s = 2 m and T_p = 9.25 s (corresponding to T_e = 8 s from spectral analysis), respectively. Sea water with a density $\rho = 1025$ kg/m^3 is considered. The gravitational acceleration is assumed to be $g = 9.81$ m/s^2.

2.3.1 Point absorber with single degree of freedom

First of all, a simple point absorber that comprises a float reacting against the sea floor is considered, as shown in Figure 2.7a and Figure 2.7b for a hydraulic and an electromechanical PTO system, respectively. With the former system, the mechanical energy derived from the motions of the float due to the wave excitation with respect to the sea floor is converted into hydraulic and then electrical energy by the PTO system. With the latter system, the mechanical energy is converted directly into electrical energy. For simplicity, the body is assumed to be constrained to motions in heave.

The corresponding free-body diagram can be seen in Figure 2.8.

2.3.1.1 Hydrodynamic coefficients

The hydrodynamic coefficients for the selected floater geometry have been calculated using the commercial software WAMIT (2013) by employing the analytical geometry and higher-order methods functions. The non-dimensional coefficients are shown in Figure 2.9. Note that the non-dimensional coefficients in heave and frequency are calculated as follows (Falnes, 2005):

$$A^*_{3,3}(\omega) = \frac{A_{3,3}}{\rho\nabla} \tag{2.79}$$

$$B^*_{3,3}(\omega) = \frac{B_{3,3}}{\rho\nabla} \tag{2.80}$$

$$f^*_3(\omega) = \frac{f_{\mathrm{e},3}}{\rho g \zeta_{\mathrm{a}} \nabla^{2/3}} \tag{2.81}$$

$$\omega^* = \omega\sqrt{\frac{r}{g}} \tag{2.82}$$

where r and ∇ are the radius and displaced volume of the cylinder, respectively. Additionally, the magnitude of the frequency-domain response amplitude operator obtained from (2.53) is plotted in Figure 2.10 against the wave period.

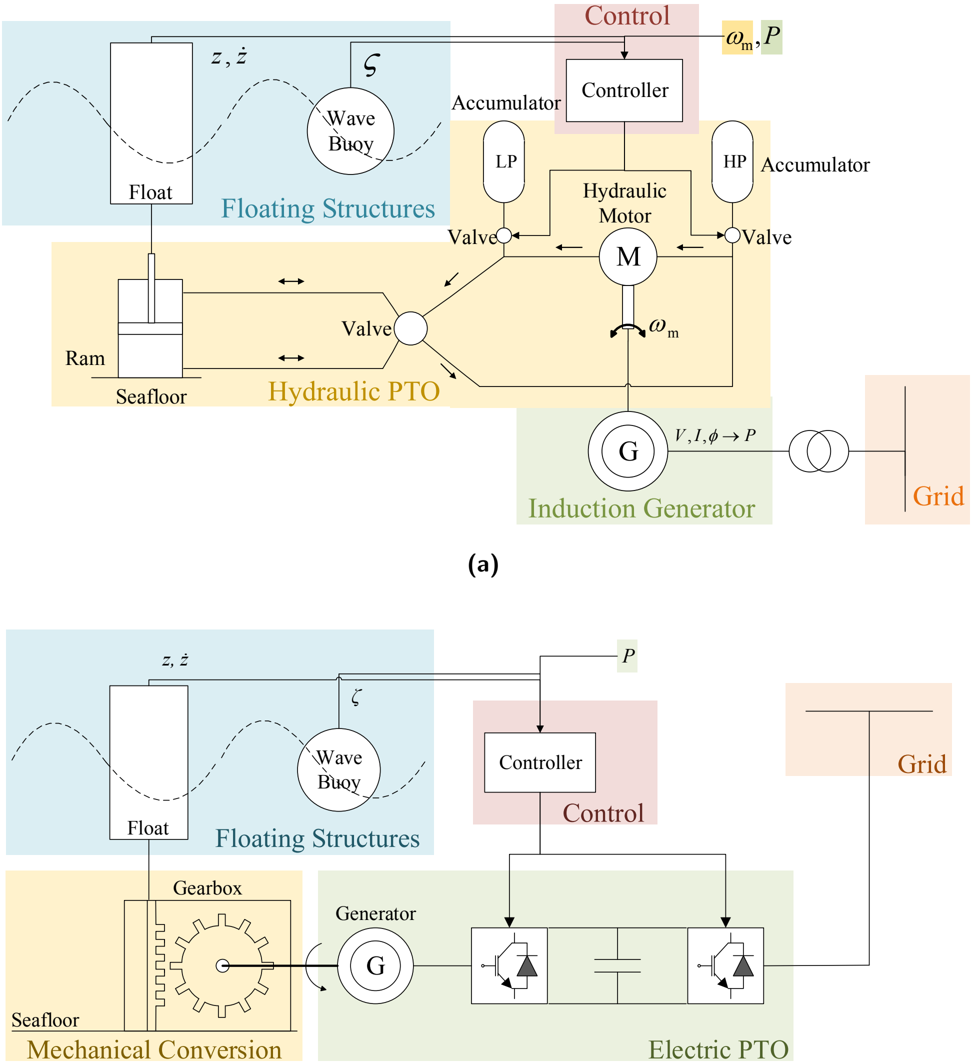
2.3.1.2 Time-domain dynamic model

Using the method described in Section 2.2.6, it is possible to obtain the following matrices for the state-space approximation of the radiation coefficients:

$$\boldsymbol{A}_{\mathrm{ss}} = \begin{bmatrix} -0.4625 & 1 & 0 & 0 \\ -0.3291 & -0.4625 & 0.3480 & 0.9764 \\ 0 & 0 & -0.5580 & 1.1738 \\ 0 & 0 & -1.1738 & -0.5580 \end{bmatrix}, \boldsymbol{B}_{\mathrm{ss}} = \begin{bmatrix} 0 \\ 134.6859 \\ 0 \\ 126.0996 \end{bmatrix}, \tag{2.83a}$$

$$\boldsymbol{C}_{\mathrm{ss}} = \begin{bmatrix} -61.5541 & 129.5137 & 0 & 0 \end{bmatrix}, \boldsymbol{D}_{\mathrm{ss}} = \boldsymbol{0}. \tag{2.83b}$$

These matrices present a maximum condition number of 4.6106. The added mass in heave at infinite wave frequency is given by $A_{3,3}(\infty) = 243.081$ tonnes. In order to obtain the equations of motion of the float, it is necessary to introduce the variable $w = \dot{z}$ that represents the vertical velocity of the body, with z indicating the heave displacement, as shown in Figure 2.8. The equations of motion of the float is expressed by (2.78). Since the float is constrained to motions in heave, the matrices and vectors



(a)

(b)

Figure 2.7: Diagram of the single-degree-of-freedom point absorber with a hydraulic (a) and an electromechanical (b) PTO system.

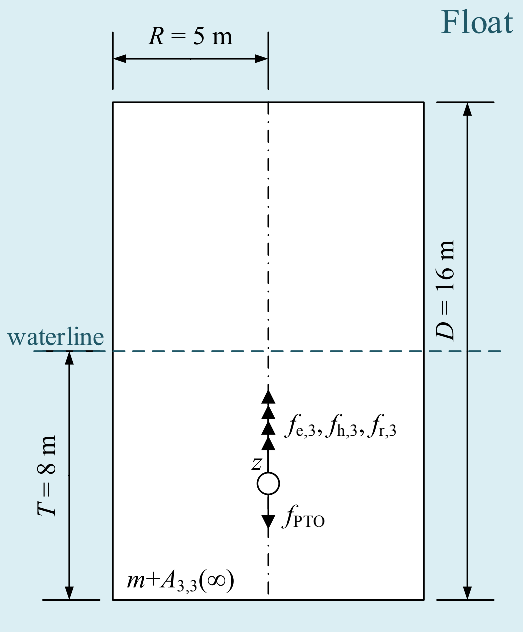


Figure 2.8: Free-body diagram and dimensions of the floating vertical cylinder point absorber with a single degree of freedom: heave.

in (2.78) reduce to:

$$\boldsymbol{x} = \begin{bmatrix} z & w & \boldsymbol{x}_{\mathrm{ss}}^T \end{bmatrix}^T, \tag{2.84a}$$

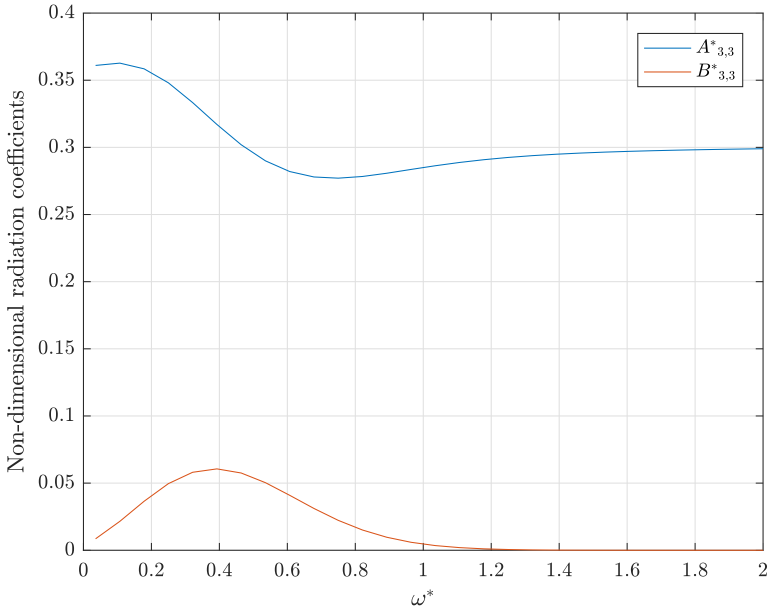
$$\boldsymbol{u} = f_{\mathrm{e},3} - f_{\mathrm{PTO}}, \tag{2.84b}$$

$$\boldsymbol{M} = m,\ \boldsymbol{A}(\infty) = A_{3,3}(\infty),\ \boldsymbol{C} = C_{3,3},\ \boldsymbol{B}_{\mathrm{m}} = \boldsymbol{0},\ \boldsymbol{C}_{\mathrm{m}} = \boldsymbol{0}. \tag{2.84c}$$

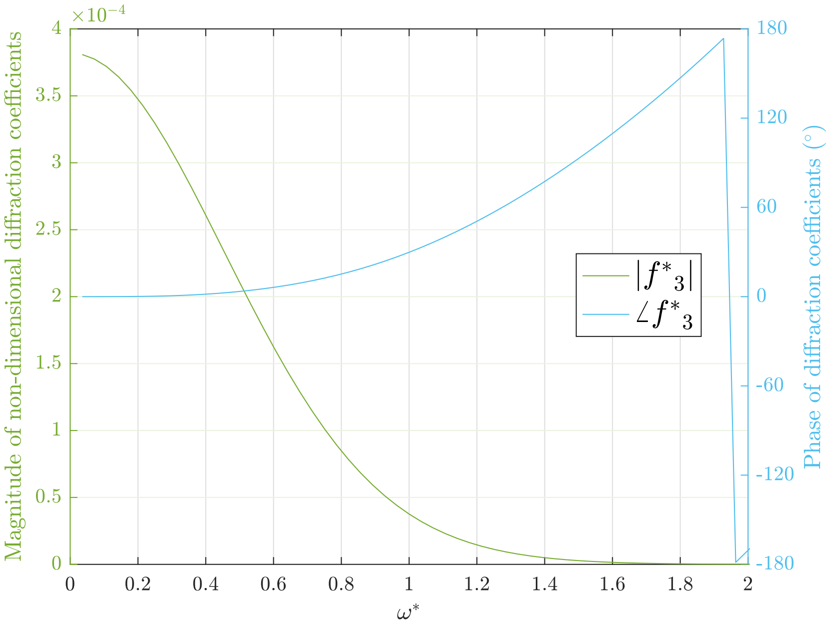
The equations of motions are discretized with a first-order-accurate Euler scheme (Süli and Mayers, 2003) and a time step of 0.1 s.

2.3.1.3 Free motions in regular and irregular waves

The motion of the point absorber with a single degree of freedom is simulated in both regular and irregular waves. For simplicity, at this stage no control force is applied by the PTO system. The wave elevation, float displacement and velocity can be seen in Figure 2.11 and Figure 2.12 for the regular and irregular wave traces, respectively. From Figure 2.10 and Figure 2.11, it is interesting to notice that the amplitude of the body displacement in the time domain is within 1% of amplitude of the response amplitude operator for the same wave period.



(a)



(b)

Figure 2.9: Variation of the non-dimensional radiation (a) and diffraction (b) coefficients with non-dimensional circular wave frequency for the floating vertical cylinder with a radius of 5 m and a draught of 8 m.

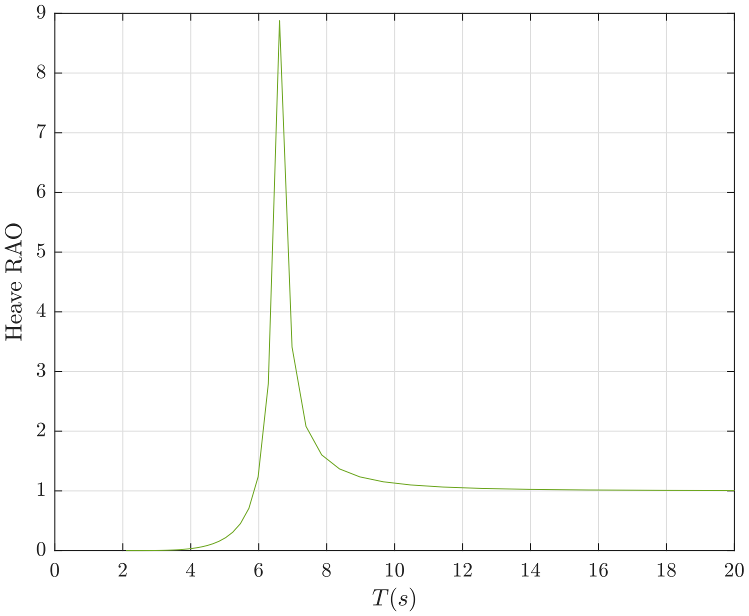


Figure 2.10: Magnitude of the response amplitude operator against wave period for the floating vertical cylinder with a radius of 5 m and a draught of 8 m.

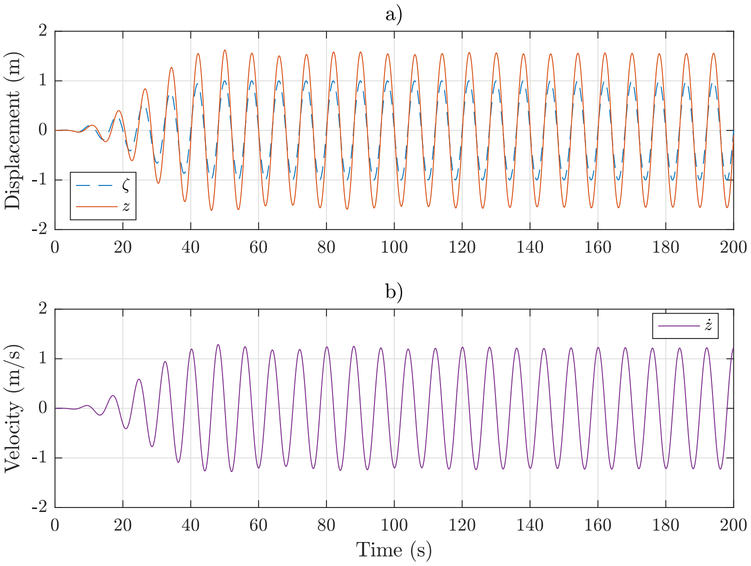


Figure 2.11: Wave elevation, float displacement (a) and velocity (b) in regular waves of unit amplitude and a period of 8 s with no control force being applied.

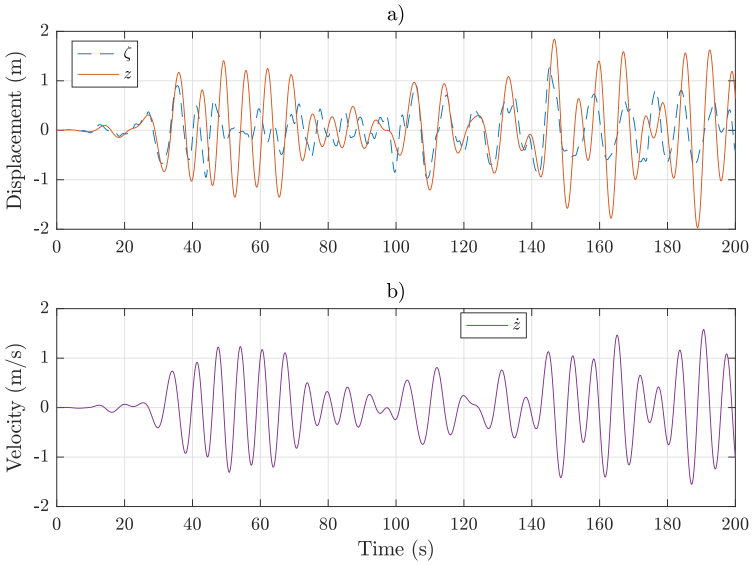


Figure 2.12: Wave elevation, float displacement (a) and velocity (b) in irregular waves with a Bretschneider spectrum with $H_\mathrm{s} = 2$ m and $T_\mathrm{p} = 9.25$ s (corresponding to $T_\mathrm{e} = 8$ s) with no control force being applied.

2.3.2 Point absorber with floater and reaction plate

The two-body point absorber analysed here is inspired by the reference model 3 (RM3) developed by the National Renewable Energy Laboratory and Sandia National Laboratories. The development of the device is described in Neary *et al.* (2014), the experimental testing of three models with different scales is reported in Yu *et al.* (2015) and validation of numerical studies is discussed in Previsic *et al.* (2014). The WEC comprises of two axisymmetric bodies: a float and a reaction plate. As shown in Figure 2.13, energy is extracted by a hydraulic PTO from the relative motion of the float with respect to the reaction plate. While the float follows the displacement of the wave elevation, the reaction plate presents motions of a much smaller magnitude due to its depth, high inertia and high viscous drag. The geometry of the point absorber can be seen in Figure 2.14, where the dimensions of the two bodies are reported. In this simplified drawing, no space is left between the inner surface of the float, which has an annular shape, and the outer surface of the spar connected to the reaction plate. In reality, there will be a gap to account for construction tolerances.

Both the float and the reaction plate are axisymmetric. Hence, within the framework of linear motions, heave is decoupled from the other degrees of freedom, although the heave degrees of freedom of the two bodies are coupled. For simplicity, in this work only heaving motions are considered, assuming the other motions to be negligible. The free body diagram of the two bodies is displayed in Figure 2.15. The float is indicated as body 1 with mass $m_1 = 727$ tonnes, while the reaction plate as body 2 with mass

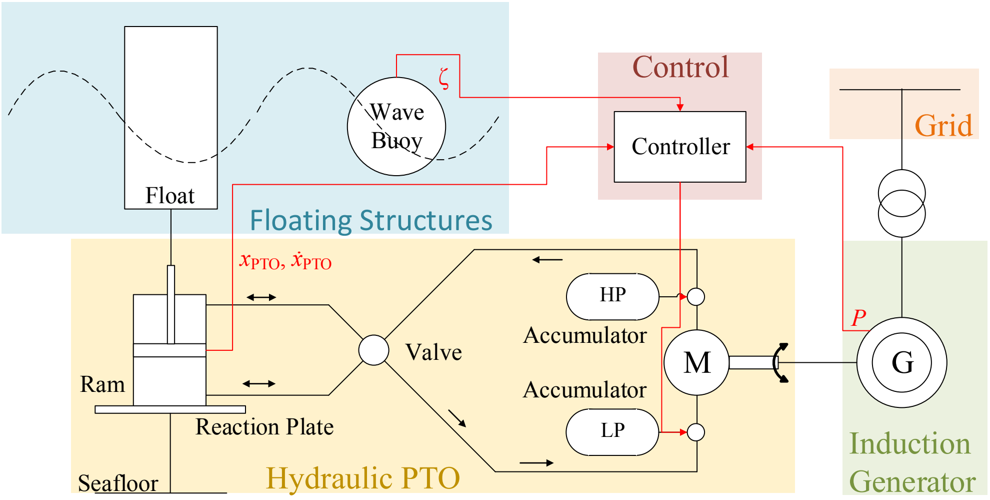


Figure 2.13: Diagram of the two-body point absorber with a hydraulic PTO system.

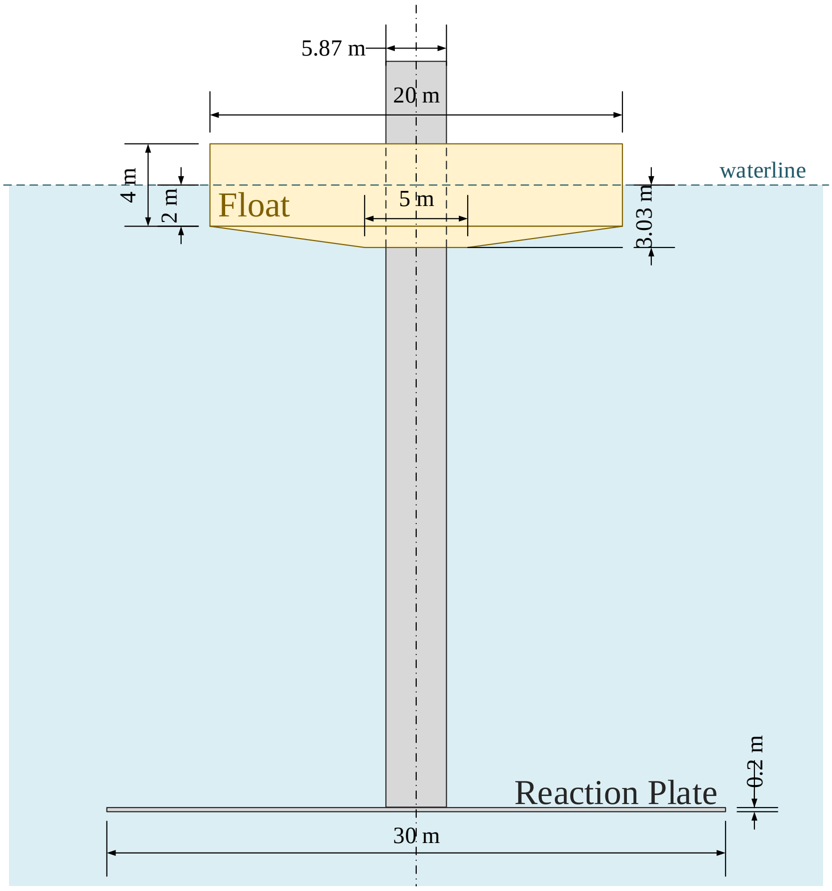


Figure 2.14: Geometry and dimensions of the float and reaction plate.

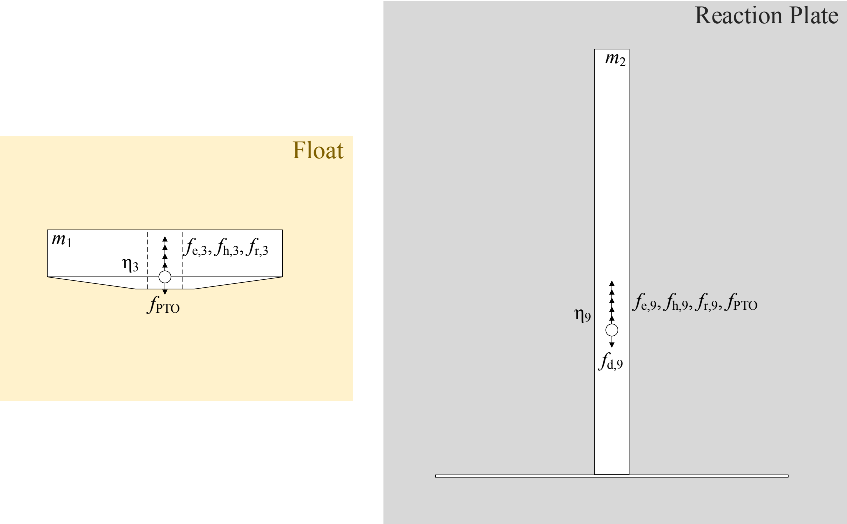

Float
m_1
η_3
$f_{e,3}, f_{h,3}, f_{r,3}$
f_{PTO}
Reaction Plate
m_2
η_9
$f_{e,9}, f_{h,9}, f_{r,9}, f_{PTO}$
$f_{d,9}$

Figure 2.15: Free-body diagram of the float and reaction plate constrained to heaving motions.

$m_2 = 912.7$ tonnes. The heave degree of freedom of the float is thus described by the digit 3, while the heave degree of freedom of the reaction plate by the digit 9, with the corresponding displacements being η_3 and η_9, respectively. The float is not connected to any moorings, so that its heaving restoring stiffness is $C_{3,3} = 2.8868$ MN/m (2.42) and (2.45). The reaction plate presents a very small water-plane area, which corresponds to the cross-section of the spar. However, a mooring system is envisioned so that the overall stiffness coefficient of the second body is assumed to be $C_{9,9} = C_{9,9}+C_{\mathrm{m},9,9} = 10$ MN/m. Therefore, the reduced inertia and stiffness matrices are, respectively,

$$\boldsymbol{M} = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \text{ and } \boldsymbol{C} = \begin{bmatrix} C_{3,3} & 0 \\ 0 & C_{9,9} \end{bmatrix}. \tag{2.85}$$

Furthermore, the damping force of the mooring system is neglected, i.e. $\boldsymbol{B}_\mathrm{m} = \mathbf{0}$.

In addition, while the viscous drag force of the float is ignored, a constant drag coefficient $C_\mathrm{d} = 5$ is applied to the reaction plate, which has been measured experimentally in Previsic *et al.* (2014). The characteristic area is set to $A_\mathrm{d} = \pi\dot{1}5^2$ m^2. The viscous drag force on the reaction plate is the computed from (2.76).

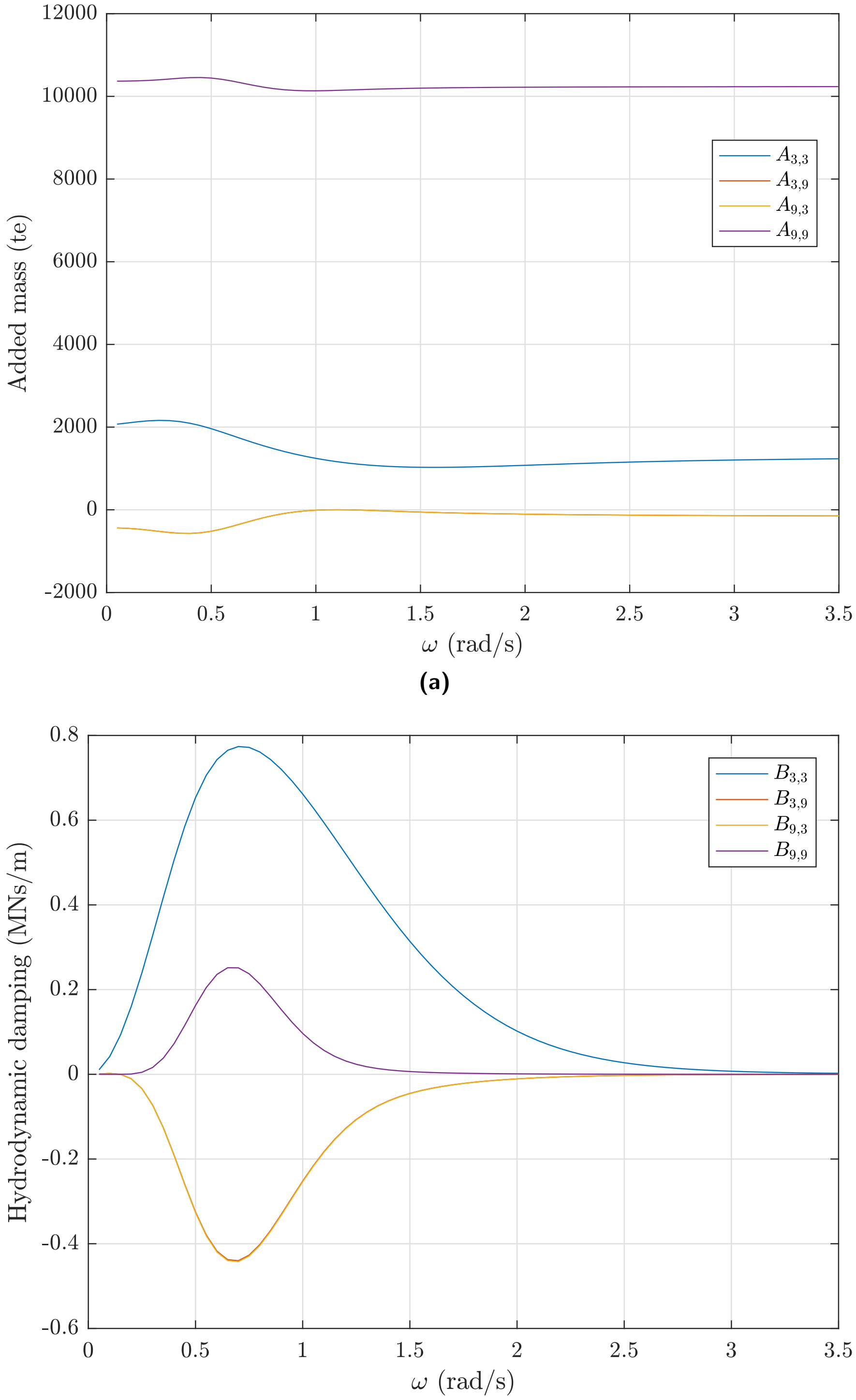
2.3.2.1 Hydrodynamic coefficients

The hydrodynamic coefficients for the selected floater geometry have been calculated using the commercial software WAMIT (2013). The float and the reaction plate have been modelled as two distinct bodies (rather than developing generalized modes), but their interactions have been accounted for. Due to their complex geometries, they have been discretized with a number of panels of sufficient quality and the low-order simulation method has been used. Care has been taken in ensuring the points on the panels on the inner surface of the float match those on the outer surface of the spar. Furthermore, the thin reaction plate has been modelled with special dipole panels (WAMIT, 2013). In Figure 2.16a, 2.16b and 2.17, it is possible to see the variation of the computed added mass, hydrodynamic damping and the wave excitation force with circular wave frequency, respectively.

2.3.2.2 Time-domain dynamic model

Employing the approach described in Section 2.2.6 developed by Forehand *et al.* (2016), it is possible to obtain the following matrices for the state-space approximation of the radiation coefficients:

$$\boldsymbol{A}_{\mathrm{ss}} = \begin{bmatrix} \boldsymbol{A}_{\mathrm{ss},1} & \boldsymbol{A}_{\mathrm{ss},2} & \boldsymbol{A}_{\mathrm{ss},3} & \boldsymbol{A}_{\mathrm{ss},4} \end{bmatrix} \text{ where} \tag{2.86}$$



(a)

(b)

Figure 2.16: Variation of the added mass (a) and hydrodynamic damping (b) with circular wave frequency for the heaving motions of the floater (3) and the spar plate (9) of the RM3 WEC.

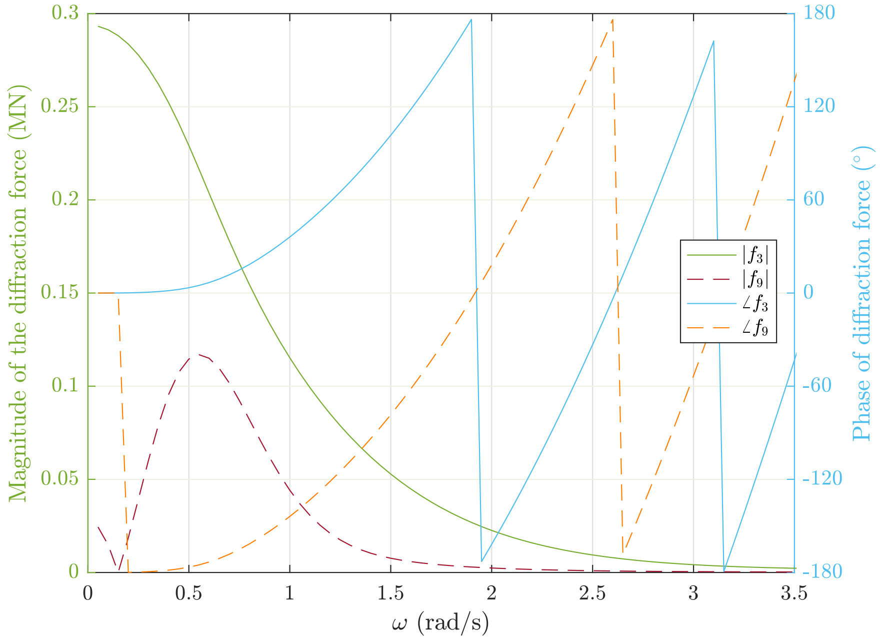


Figure 2.17: Variation of the wave excitation force with circular wave frequency for the heaving motions of the floater (3) and the spar plate (9) of the RM3 WEC.

$$
\boldsymbol{A}_{\mathrm{ss},1} = \begin{bmatrix}
-0.4684 & 1 & 0 & 0 & 0 \\
-0.1401 & -0.4684 & 6.2358 & 2.3048 & -1327943.7066 \\
0 & 0 & -1.0458 & 1.3292 & 0 \\
0 & 0 & -1.3292 & -1.0458 & -7996171.4699 \\
0 & 0 & 0 & 0 & -1597207958577.22 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}, \qquad (2.87)
$$

$$\boldsymbol{A}_{\mathrm{ss},2} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -0.4069 & 1 & 0 & 0 & 0 \\ -0.1494 & -0.4069 & -0.8309 & 0.8143 & 0 \\ 0 & 0 & -0.5234 & 1 & 0 \\ 0 & 0 & -0.5923 & -0.5234 & 0 \\ 0 & 0 & 0 & 0 & -0.3290 \\ 0 & 0 & 0 & 0 & -0.1733 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.88}$$

$$\boldsymbol{A}_{\mathrm{ss},3} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ -0.3290 & -0.8071 & 0.5003 & 0.2673 & -0.3447 \\ 0 & -0.4779 & 1 & 0 & 0 \\ 0 & -0.6671 & -0.4779 & 0.2414 & -0.3112 \\ 0 & 0 & 0 & -0.7220 & 3.3249 \\ 0 & 0 & 0 & -3.3249 & -0.7220 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.89}$$

$$\boldsymbol{A}_{\mathrm{ss},4} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -0.2954 & 1 & 0 & 0 & 0 & 0 \\ -0.2533 & -0.2954 & -0.0571 & 1.0196 & 0 & 0 \\ 0 & 0 & -0.3992 & 1 & 0 & 0 \\ 0 & 0 & -0.71902 & -0.3992 & 0.3955 & -4.4659 \\ 0 & 0 & 0 & 0 & -1.9059 & 1 \\ 0 & 0 & 0 & 0 & 0 & -20.3028 \end{bmatrix}, \qquad (2.90)$$

$$\boldsymbol{B}_{\mathrm{ss}} = \begin{bmatrix} 0 & 0 \\ 275447.5122 & 0 \\ 0 & 0 \\ 1658598.5743 & 0 \\ 331299404092.387 & 0 \\ 0 & 0 \\ 265.7784 & 0 \\ 0 & 0 \\ 286.8697 & 0 \\ 0 & 0 \\ 0 & 262.5500 \\ 0 & 0 \\ 0 & 237.0285 \\ 0 & 0 \\ 0 & 108.8952 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 512.4006 \\ 0 & 0 \\ 0 & 2293.7115 \end{bmatrix}, \tag{2.91}$$

$$\boldsymbol{C}_{\mathrm{ss}} = \begin{bmatrix} \boldsymbol{C}_{\mathrm{ss},1} & \boldsymbol{C}_{\mathrm{ss},2} & \boldsymbol{C}_{\mathrm{ss},3} & \boldsymbol{C}_{\mathrm{ss},4} \end{bmatrix} \text{ where} \tag{2.92a}$$

$$\boldsymbol{C}_{\mathrm{ss},1} = \begin{bmatrix} -74616019813.8840 & 159452826352.593 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.92b}$$

$$\boldsymbol{C}_{\mathrm{ss},2} = \begin{bmatrix} 0 & 0 & 0 & 0 & 212.8615 \\ 243.4395 & -605.2666 & 0 & 0 & 0 \end{bmatrix}, \tag{2.92c}$$

$$\boldsymbol{C}_{\mathrm{ss},3} = \begin{bmatrix} -652.7569 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.92d}$$

$$\boldsymbol{C}_{\mathrm{ss},4} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -889.7882 & 3065.6494 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.92e}$$

$$\boldsymbol{D}_{\mathrm{ss}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{2.93}$$

It is important to notice that the size of the matrices is due to the approximation system modelling also the coupling between the radiation force of the floater and reaction plate.

The added mass matrix at infinite wave frequency is given by

$$\boldsymbol{A}(\infty) = \begin{bmatrix} 1284.858 & -162.148 \\ -161.807 & 10241.712 \end{bmatrix} \text{ tonnes.} \tag{2.94}$$

For this particular WEC, the time-domain equations of motions is given by (2.78) with

$$\boldsymbol{x} = \begin{bmatrix} \eta_3 & \eta_9 & \nu_3 & \nu_9 & \boldsymbol{x}_{\mathrm{ss}}^T \end{bmatrix}^T, \tag{2.95a}$$

$$\boldsymbol{f}_{\mathrm{e}} = \begin{bmatrix} f_{\mathrm{e},3} & f_{\mathrm{e},3} \end{bmatrix}^T, \tag{2.95b}$$

$$\boldsymbol{f}_{\mathrm{PTO}} = \begin{bmatrix} -f_{\mathrm{PTO}} & f_{\mathrm{PTO}} \end{bmatrix}^T, \tag{2.95c}$$

$$\boldsymbol{f}_{\mathrm{d}} = \begin{bmatrix} 0 & f_{\mathrm{d},9} \end{bmatrix}^T. \tag{2.95d}$$

The equations of motions are discretized with a first-order-accurate Euler scheme (Süli and Mayers, 2003) and a time step of 0.1 s.

2.3.2.3 Free motions in regular and irregular waves

The motion of the float and reaction plate is simulated in both regular and irregular waves. For simplicity, at this stage no control force is applied by the PTO system. The wave elevation, float and reaction plate heave displacement and velocity can be seen in Figure 2.18 and Figure 2.19 for the regular and irregular wave traces, respectively.

As can be seen in Figure 2.18, in regular waves the heave displacement and velocity of the reaction plate is much smaller in magnitude than those of the float, as expected. Whereas the float is almost in phase with the wave elevation, the reaction plate is almost completely out of phase. In irregular waves (Figure 2.19), the float still follows the wave elevation closely. Conversely, the behaviour of the reaction plate is more difficult to understand, despite the much lower magnitude of its motion.

2.3.3 Point absorber with direct-drive PTO

The team at Uppsala University has developed over the years a sea-floor-referenced point absorber with a direct-drive PTO system. The development and testing of a number of full-scale prototypes of this device, known as Sebased, is well described in the literature (Danielsson, 2006; Eriksson, 2007; Waters, 2008; Stalberg *et al.*, 2008; Lejerskog *et al.*, 2015).

Figure 2.20 shows a diagram of the device, which is inspired by Eriksson *et al.* (2007). A small float, excited by incident waves, drives a linear, permanent-magnet generator

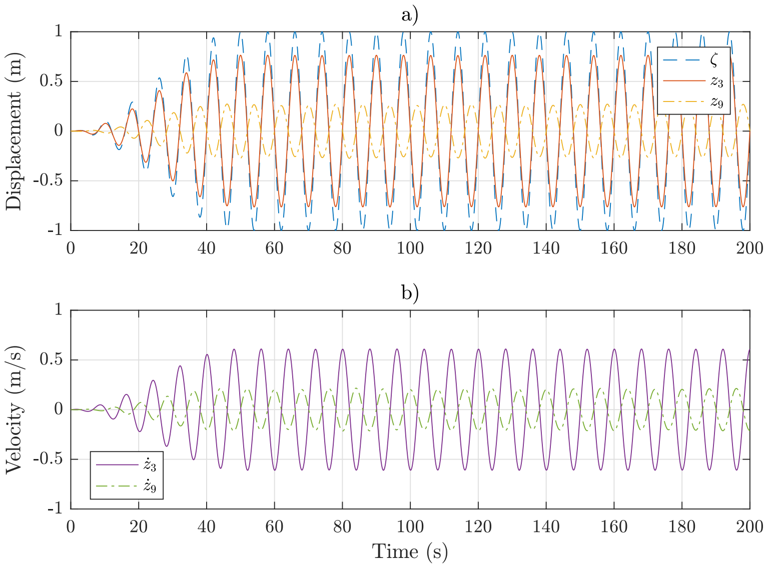


Figure 2.18: Wave elevation, heave displacement (a) and velocity (b) of the float (3) and reaction plate (9) in regular waves of unit amplitude and a period of 8 s with no control force being applied.

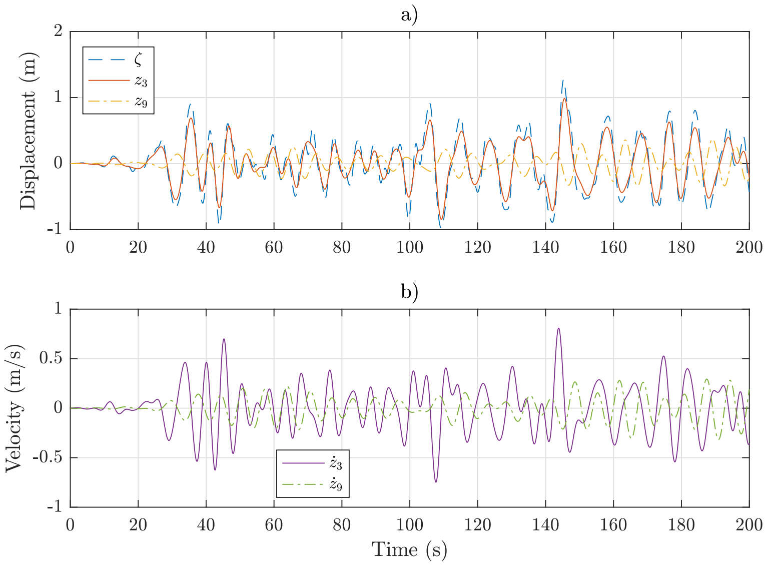


Figure 2.19: Wave elevation, heave displacement (a) and velocity (b) of the float (3) and reaction plate (9) in irregular waves with a Bretschneider spectrum with $H_\mathrm{s} = 2$ m and $T_\mathrm{p} = 9.25$ s (corresponding to $T_\mathrm{e} = 8$ s) with no control force being applied.

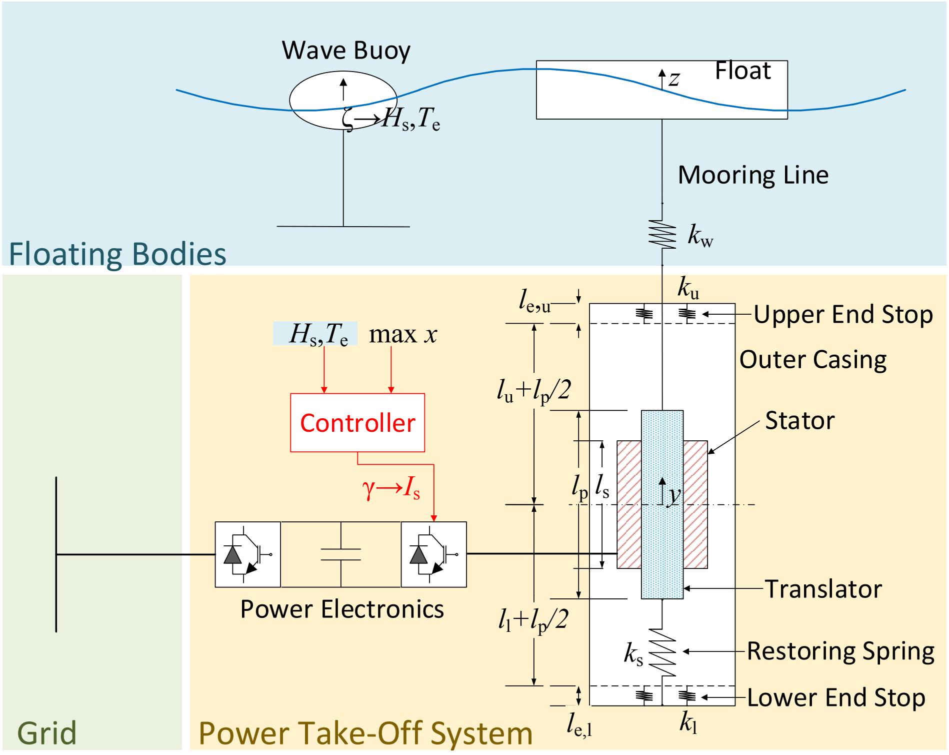


Figure 2.20: Diagram of the prototype Seabased WEC.

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 upper and lower end stops prevent the translator from breaking the casing in high waves. The motion of the magnet induces electrical current in the coils wound around the stator. Power absorption is controlled through 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 (Waters, 2008).

In Figure 2.20, the same naming convention as in Eriksson *et al.* (2007) is held, with the values of the variables quantities being given in Table 2.2. The importance of the nominal rating, voltage and velocity of the generator is explained in the next chapter. 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 Waters (2008).

The float of the first Seabased prototype is a vertical cylinder with a radius of 1.5 m, a height of 0.8 m, a draught of 0.4 m and a mass of 1000 kg. Furthermore, in this work a study will be performed assuming a change in the geometry due to marine bio-fouling.

Table 2.2: Main features of the Seabased WEC, taken from Eriksson *et al.* (2007).

f_0 (kN)	8.12	P_{nom} (kW) of generator	10
V_{nom} (V)	133	v_{nom} (m/s)	0.67
Pole width (m)	0.050	Piston mass (kg)	1000
l_{p} (m)	1.867	l_{s} (m)	1.264
k_{u} (kN/m)	243	k_{l} (kN/m)	215
k_{w} (kN/m)	450	k_{s} (kN/m)	6.2

As a result, the buoy geometry is assumed to change to a vertical cylinder with a radius of 1.75 m and a draught of 0.5 m. The corresponding increase in mass is assumed to be equal to the change in displaced volume. Hence, two float geometries are considered.

2.3.3.1 Hydrodynamic coefficients

The hydrodynamic coefficients for the selected floater geometries have been calculated using the commercial software WAMIT (2013) employing an analytical geometry and the higher order method. The non-dimensional coefficients are shown in Figure 2.21 for both the original and modified floater. Furthermore, the magnitude of the frequency-domain response amplitude operator obtained from (2.53) is plotted against the wave period for the two geometries in Figure 2.22.

2.3.3.2 Time-domain dynamic model

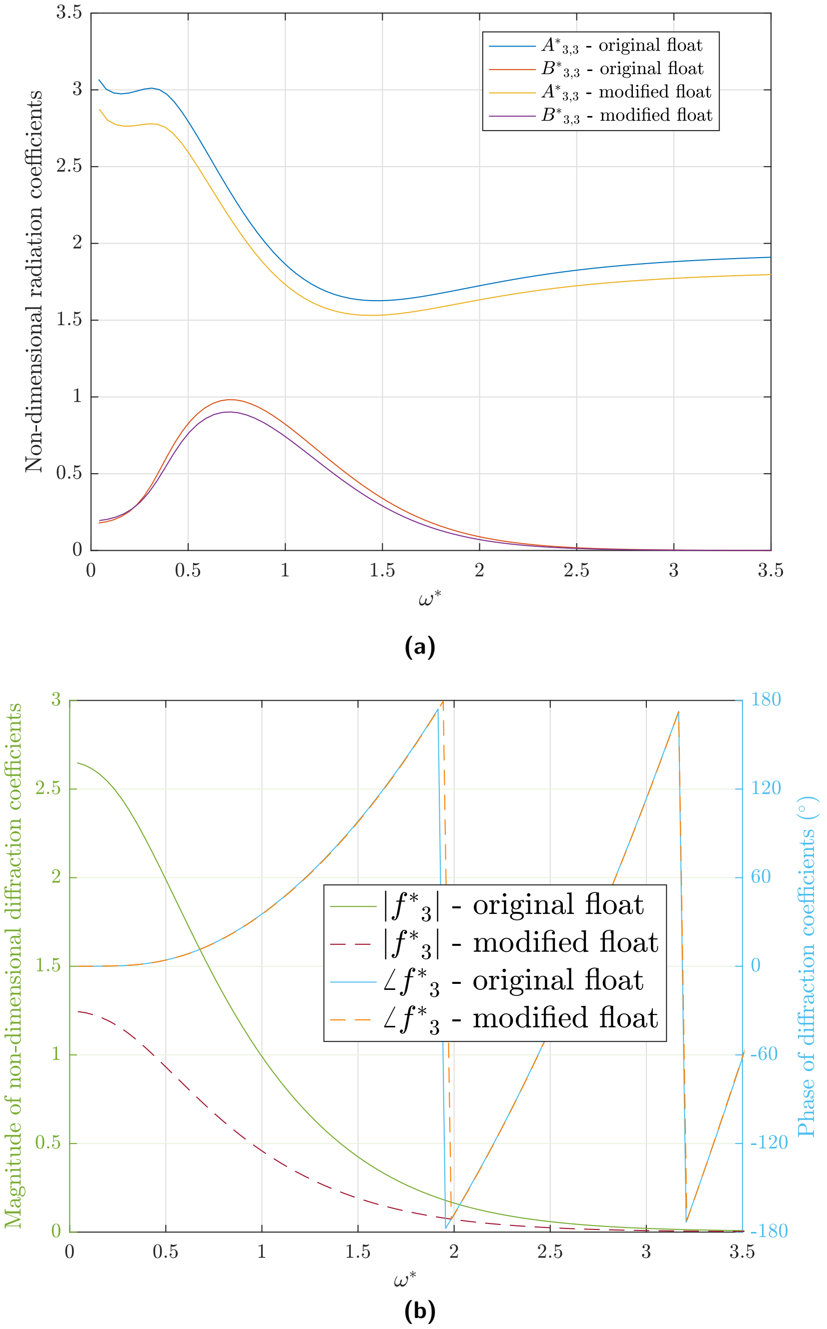
A weakly non-linear mathematical model of the system dynamics has been developed by Eriksson *et al.* (2007). As this section is adapted from that work, the reference is no longer cited in this section.

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 considered negligible (Eriksson *et al.*, 2006). This reduces the system to two degrees of freedom: the displacements of the float and translator, which are labelled as z and y, respectively. From the analysis the free-body diagram shown in Figure 2.23, the motions of the two bodies are expressed through the following system of equations (Eriksson *et al.*, 2007)

$$(m_{\mathrm{b}} + A_{3,3}(\infty))\,\ddot{z}(t) = f_{\mathrm{e}}(t) - f_{\mathrm{r}}(t) - f_{\mathrm{h}}(t) - f_{\mathrm{w}}(t), \tag{2.96a}$$

$$m_{\mathrm{p}}\ddot{y}(t) = f_{\mathrm{m}}(t) - f_{\mathrm{PTO}}(t) - f_{\mathrm{s}}(t) + f_{\mathrm{u}}(t) + f_{\mathrm{l}}(t), \tag{2.96b}$$

where m_{b} and m_{p} the mass of the float and piston respectively, f_{w} the tension in the wire connecting the float to the translator, f_{s} the force of the restoring spring in Figure 2.20, 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},



(a)

(b)

Figure 2.21: Variation of the non-dimensional radiation (a) and diffraction (b) coefficients with non-dimensional circular wave frequency for the two floating vertical cylinders with a radius of 1.5 m and a draught of 0.4 m (original) and 1.75 m and 0.5 m (modified), respectively.

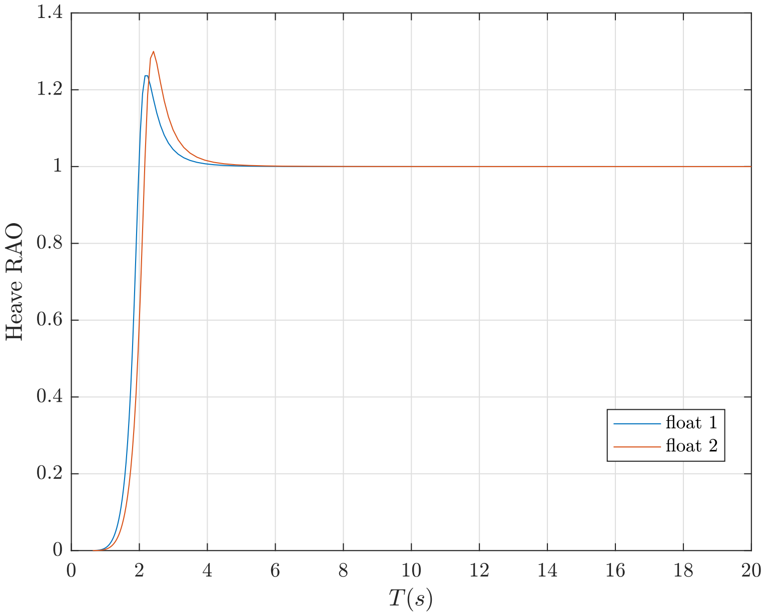


Figure 2.22: Magnitude of the response amplitude operator against wave period for the two floating vertical cylinders with a radius of 1.5 m and a draught of 0.4 m (original) and 1.75 m and 0.5 m (modified), respectively.

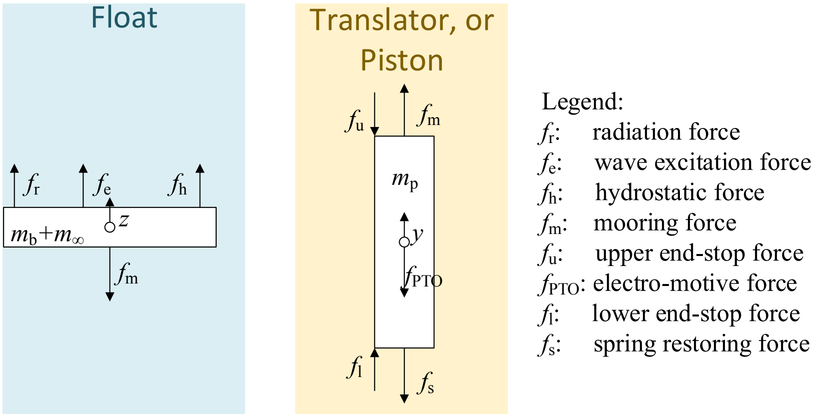


Figure 2.23: Free-body diagram of the Seabased WEC subject to heaving motions only.

which are activated only if the end stops are reached, and f_{em}, which depends on the exposure of the translator to the stator.

Using the method described in Section 2.2.6, it is possible to obtain the following matrices for the state-space approximation of the radiation coefficients for the original and modified float, respectively:

$$\boldsymbol{A}_{\mathrm{ss}} = \begin{bmatrix} -1.1538 & 1.2190 & 0 & 0 \\ -1.2190 & -1.1538 & -1.8303 & 1.9300 \\ 0 & 0 & -2.1212 & 3.2847 \\ 0 & 0 & -3.2847 & -2.1212 \end{bmatrix}, \boldsymbol{B}_{\mathrm{ss}} = \begin{bmatrix} 0 \\ 74.9326 \\ 0 \\ 145.3925 \end{bmatrix}, \tag{2.97a}$$

$$\boldsymbol{C}_{\mathrm{ss}} = \begin{bmatrix} -134.8297 & 143.7977 & 0 & 0 \end{bmatrix}, \boldsymbol{D}_{\mathrm{ss}} = \mathbf{0} \text{ and} \tag{2.97b}$$

$$\boldsymbol{A}_{\mathrm{ss}} = \begin{bmatrix} -1.0629 & 1.1919 & 0 & 0 \\ -1.1919 & -1.0629 & -1.5449 & 1.8056 \\ 0 & 0 & -1.8761 & 3.0370 \\ 0 & 0 & -3.0370 & -1.8761 \end{bmatrix}, \boldsymbol{B}_{\mathrm{ss}} = \begin{bmatrix} 0 \\ 73.7882 \\ 0131.9090 \end{bmatrix}, \tag{2.98a}$$

$$\boldsymbol{C}_{\mathrm{ss}} = \begin{bmatrix} -166.4568 & 190.0697 & 0 & 0 \end{bmatrix}, \boldsymbol{D}_{\mathrm{ss}} = \mathbf{0}. \tag{2.98b}$$

These matrices present a maximum condition number of 3.5556 and 3.3770, respectively. The added mass in heave at infinite wave frequency is given by $A_{3,3}(\infty) = 5.715$ tonnes and $A_{3,3}(\infty) = 9.130$ tonnes, respectively.

The electromotive, or PTO control, force f_{PTO} is discussed in the next chapter. The hydrostatic force is calculated from (2.42) and (2.45). Hence, $C_{3,3} = 71.076$ kN/m and $C_{3,3} = 96.743$ kN/m for original and modified floaters, respectively. Ignoring compression effects, the non-linear mooring force is given by

$$f_{\mathrm{m}} = \begin{cases} -k_{\mathrm{w}}(z-y) & \text{if } z > y, \qquad (2.99\mathrm{a}) \\ 0 & \text{otherwise,} \qquad (2.99\mathrm{b}) \end{cases}$$

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

$$f_{\mathrm{u}} = \begin{cases} -k_{\mathrm{u}}(y-l_{\mathrm{u}}) & \text{if } y > l_{\mathrm{u}}, \qquad (2.100\mathrm{a}) \\ 0 & \text{otherwise,} \qquad (2.100\mathrm{b}) \end{cases}$$

$$f_{\mathrm{l}} = \begin{cases} -k_{\mathrm{l}}(y+l_{\mathrm{l}}) & \text{if } y < -l_{\mathrm{l}}, \qquad (2.101\mathrm{a}) \\ 0 & \text{otherwise,} \qquad (2.101\mathrm{b}) \end{cases}$$

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 Figure 2.20. The force of the spring connected to the translator is expressed as

$$f_\mathrm{s} = f_0 + k_\mathrm{s} y, \tag{2.102}$$

where f_0 is a static force due to pre-charging, and k_s is the spring stiffness. The values of l_l, l_u, k_l, k_u, k_w, k_s, m_b, m_p, f_0 and S_w for the Seabased device can be seen in Table 2.2.

Using (2.99-2.102), (2.96) has been expressed in the following non-linear state-space form

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t,x) + \boldsymbol{B}\boldsymbol{w}(t) + \boldsymbol{B}\boldsymbol{l}(t,z,y), \text{ where} \tag{2.103a}$$

$$\boldsymbol{x} = \begin{bmatrix} z(t) & \dot{z}(t) & y(t) & \dot{y}(t) & \boldsymbol{x}_\mathrm{ss}^T(t) \end{bmatrix}^T, \tag{2.103b}$$

$$\boldsymbol{u} = \begin{bmatrix} 0 & -f_\mathrm{PTO}(x) \end{bmatrix}^T, \tag{2.103c}$$

$$\boldsymbol{w} = \begin{bmatrix} f_\mathrm{e}(t) & 0 \end{bmatrix}^T, \tag{2.103d}$$

$$\boldsymbol{l} = \begin{bmatrix} -f_\mathrm{m}(x,y) & f_\mathrm{m}(x,y) - f_0 + f_\mathrm{u}(x) + f_\mathrm{l}(x) \end{bmatrix}^T, \tag{2.103e}$$

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \boldsymbol{0}^T \\ -C_{3,3}/\left(m_\mathrm{b} + A_{3,3}(\infty)\right) & 0 & 0 & 0 & -\boldsymbol{C}_\mathrm{ss}/\left(m_\mathrm{b} + A_{3,3}(\infty)\right) \\ 0 & 0 & 0 & 1 & \boldsymbol{0}^T \\ 0 & 0 & -k_\mathrm{s}/m_\mathrm{p} & 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{B}_\mathrm{ss} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_\mathrm{ss} \end{bmatrix}, \tag{2.103f}$$

$$\boldsymbol{B} = \begin{bmatrix} 0 & \left(m_\mathrm{b} + A_{3,3}(\infty)\right)^{-1} & 0 & 0 & \boldsymbol{0}^T \\ 0 & 0 & 0 & 1/m_\mathrm{p} & \boldsymbol{0}^T \end{bmatrix}^T. \tag{2.103g}$$

The equations of motions are discretized with a fourth-order-accurate Runge-Kutta scheme (Süli and Mayers, 2003) and a time step of 0.01 s.

2.3.3.3 Motions in regular and irregular waves

The motion of the Seabased device is also simulated in the regular and irregular wave traces. However, in this case it is necessary to specify a control force due to instabilities in the numerical model for the case of no electromotive force. Here, the PTO force is specified to be proportional to the vertical velocity of the translator, $f_\mathrm{PTO} = B_\mathrm{PTO}\dot{y}$, with $B_\mathrm{PTO} = 10$ kNs/m. This value is the smallest possible value to ensure stability. The model of the PTO system and the control strategy are described in detail in the next chapter.

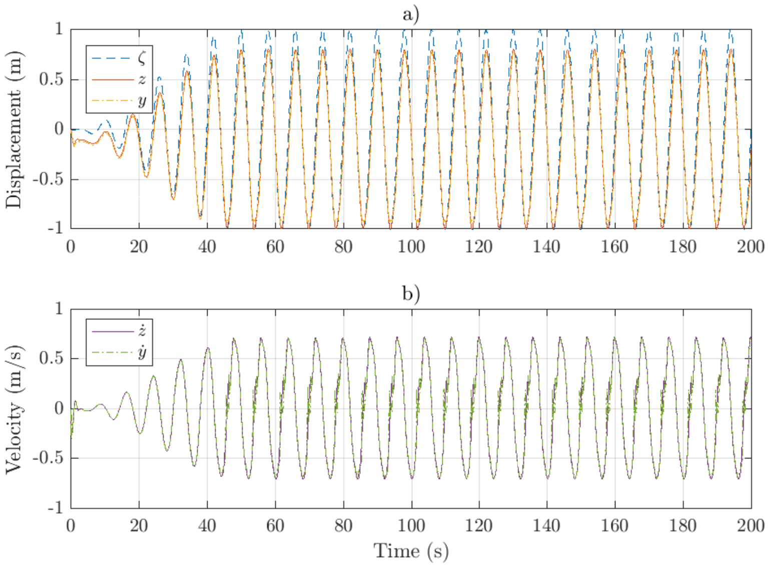


Figure 2.24: Wave elevation, heave displacement (a) and velocity (b) of the float (z) and translator (y) in regular waves of unit amplitude and a period of 8 s with no control force being applied.

For simplicity, only the motions associated with the original float are displayed here. The resulting displacement and velocity of the buoy and translator can be seen in Figure 2.24 and Figure 2.25 for the regular and irregular wave traces, respectively.

As can be seen in Figure 2.24 and Figure 2.25, in both regular and irregular waves both the float and translator follows closely the wave elevation due to the high mooring stiffness and the ineffectiveness of the end stops in these mild wave conditions. The velocity of the translator presents some instabilities due to the bang-bang nature of the mooring force. A more realistic model could be obtained with a different formulation of the mooring force and a finer time step.

2.4 Chapter summary

In this chapter, after explaining briefly the working mechanisms of WECs and their PTO units, methods for the modelling of their dynamics have been discussed. In particular, potential flow theory with the addition of some non-linear effects has been selected to be used in this thesis due to its good compromise of accuracy and computational cost. Subsequently, the theory of wave-body interactions has been described in detail and the equations of motion of a WEC have been derived in both the frequency and time domains. Finally, three different point absorber technologies with increasing complexity have been introduced and modelled. These models will be used throughout this document for the testing and assessment of the developed control strategies.

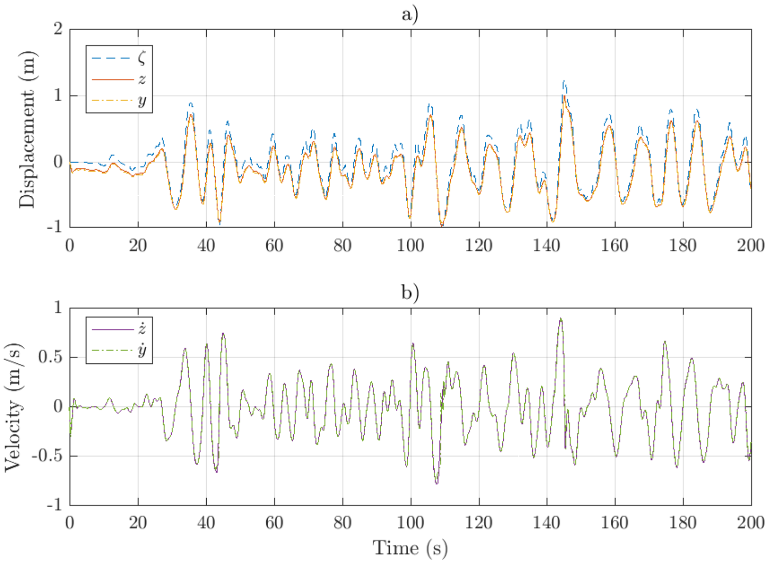


Figure 2.25: Wave elevation, heave displacement (a) and velocity (b) of the float (z) and translator (y) in irregular waves with a Bretschneider spectrum with $H_\mathrm{s} = 2$ m and $T_\mathrm{p} = 9.25$ s (corresponding to $T_\mathrm{e} = 8$ s) with no control force being applied.

Chapter 3

Control of wave energy converters: state-of-the-art

In this chapter, the state-of-the-art schemes for WEC control are discussed. After providing a background summary of existing technologies, some strategies will be analysed in detail, as they will form the basis of the learning algorithms developed in this thesis.

3.1 Literature review

Since the 1970s, multiple control strategies have been proposed for the maximization of energy absorption of WECs. A review of the initial schemes can be found in Salter *et al.* (2002). More recent developments can be found in the review by Ringwood *et al.* (2014) and the book by Korde and Ringwood (2016), which is dedicated to the topic of WEC control. Whereas the first schemes were based purely on hydrodynamic considerations, there has been a growing technology transfer in the past decade from the field of control systems, which has resulted in the application of innovative techniques, such as model predictive control. Even greater innovation is expected as a result of the current control call by Wave Energy Scotland (2017).

The main challenge with WEC control is that the control problem is non-causal, i.e. information on both the past and current wave excitation force is required in order to select the optimal control action. This requires the prediction of the wave elevation over a future time horizon. Fusco and Ringwood (2010a) have employed cyclical models, an extended Kalman filter, autoregressive models and neural networks for the forecast of the wave elevation at a defined point in space. Autoregressive models have been found to have the best performance, with a good accuracy over a couple of wave periods in the future with swell waves. An alternative approach based on deterministic sea wave prediction was proposed by Li *et al.* (2012), which employs a network of buoys or LIDAR. Nevertheless, these approaches have not been tested in the actual environment yet. The main problems are caused by the impossibility to measure the wave elevation at the exact location of the WEC. As a result, measurement noise and local effects (e.g. wind) can affect the estimation of the current wave elevation and spreading at the WEC position, and thus any subsequent wave prediction. For these reasons, the first studies of WEC control have assumed stationary wave conditions over periods lasting 15-30 minutes, known as sea states (Holthuijsen, 2007), during which constant control settings may be used. This technique was also adopted successfully at Pelamis Wave Power on the P1 and P2 prototypes.

From hydrodynamic considerations, the simplest control strategy consists in setting the PTO force to be proportional to the velocity at the PTO (Salter *et al.*, 2002). This method is known as passive or resistive control, and it enables the control of the amplitude of the WEC response. The main advantages of this method are its simple practical implementation and the associated absence of negative power flows (i.e. power supplied to the WEC during parts of the wave cycle). This strategy can be extended to phase and amplitude control through the addition of a stiffness term (Salter *et al.*, 2002), i.e. the PTO force is given by the combination of a term proportional to the velocity and one to the displacement. This approach, known as complex-conjugate, impedance-matching or reactive control, results in optimal energy absorption from hydrodynamic considerations (Falnes, 2005). Nevertheless, this technique suffers from a number of problems. As hinted by the name, the strategy presents reactive power flows during part of the wave cycle so as to maximize the overall extracted energy. This can result in a very high peak to mean power ratio, which would require more expensive electrical machinery. Furthermore, the scheme would cause very large displacements of and loads on the WEC, so that realistic constraints should be considered for the PTO force saturation and PTO displacement to prevent damage in energetic waves. In addition, for point absorbers the optimal stiffness coefficient is likely to be negative (Falcão, 2008). Although this can be achieved in practice with power electronics, the system may become unstable if the magnitude of the control stiffness coefficient exceeds the hydrostatic and mooring stiffness of the WEC (Wave Energy Scotland, 2016). For these reasons, suboptimal control strategies have been proposed over the years instead.

It is possible to show from first principles that the optimal damping and stiffness coefficients for the maximization of energy absorption depend on the wave frequency (Salter *et al.*, 2002; Ringwood *et al.*, 2014; Korde and Ringwood, 2016). Hence, the controller parameters should be adapted to the current wave conditions. However, this is challenging in irregular waves due to their stochastic nature. The simplest approach consists of assuming stationary sea state conditions, as aforementioned. Discrete sea states are identified as determined by statistical measures for the wave period and amplitude. In particular, for wave energy, the significant wave height and the energy wave period have been identified as the preferred parameters (Cruz, 2008). The significant wave height becomes influential in the selection of the control parameters due to the displacement constraints that may be reached in high waves. Simulations are then run with numerical models to determine the optimal control coefficients in each sea state, which are stored in a look-up table. Although originally frequency-domain models were more common (Cruz, 2008), time-domain models are now used more often due to the necessity to account for realistic saturation limits on the PTO force and non-linear effects. During the operation of the WEC, the current sea state conditions are determined by near-by wave buoys and the optimal control coefficients selected from the look-up table. The approach could be further simplified by considering seasonal settings, as investigated by Valério *et al.* (2008). Despite the simplicity of the control strategies, Pelamis Wave Power Ltd. predicted an increase in absorbed energy by as much as 100% from numerical studies using the adaptive resistive control rather than a fixed damping coefficient. During tests on the full-scale P2 prototypes, the actual improvement was quantified as 37%. Even greater power absorption is predicted with reactive control, with a increase of 50% in annual absorbed energy over resistive control being reported by Nambiar *et al.* (2015) from a numerical study of an array of point absorbers. Indeed, resistive and reactive control can be easily extended to the treatment of devices with multiple degrees of freedom and wave farms. Nevertheless, WEC performance could be improved further by adapting its response on a wave by wave approach, which would also reduce the failure risks associated with rogue waves.

As a result, a number of research groups have tried to develop adaptive control schemes, which can be solved in a realistic computational time for a real-time implementation. One of the most promising strategies is simple-but-effective control, proposed by Fusco and Ringwood (2013) and discussed in Ringwood *et al.* (2014) and Korde and Ringwood (2016). The basis of this algorithm is the assumption that the wave excitation force is a narrow-banded (i.e. sinusoidal) process. Therefore, the search for the coefficients reduces to that in regular waves. An extended Kalman filter is used to find the approximate amplitude and frequency corresponding to the current wave. The displacement constraints are expressed through an equivalent velocity, and thus gain, limit from first principles. The method is found to result in a relative capture width within 10% of model predictive control, which will be discussed later on, and even presents superior performance for long wave periods (Ringwood *et al.*, 2014), despite the lower computational cost. A similar strategy based on a Lyupanov function has been developed by Korde *et al.* (2016), although it presents a worse performance. An alternative approach that does not require wave prediction (hence, it assumes the WEC control problem to be causal) has been proposed by Zou *et al.* (2017). The method switches between circular arc and bang-bang modes based on Pontryagin's principle, presenting a performance similar to complex-conjugate control. Scruggs and Nie (2015) investigate an alternative adaptive technique that includes the prediction of the wave excitation force using a moving window approach with a Gauss-Markov innovations model. An even different approach is adopted by Cantarellas *et al.* (2017) based on an adaptive vectorial approach, which focuses on the reduction of large power fluctuations.

A completely different method for the real-time, phase control of WECs is characterized by discrete control schemes, such as latching and declutching control. Latching control has been first proposed by Budal and Falnes (1977) for the maximization of the energy absorption of point absorbers in waves with a period greater than the natural frequency of the device. In particular, the heaving motion only has been analysed. The authors showed that one condition for the maximization of the energy absorption is for the body velocity to be in phase with the excitation force. Latching control achieves this situation by locking, or latching, the body in place at the instant when its velocity vanishes at the end of one oscillation. The device is then released after a predefined time interval, which needs to be determined. The optimal solution is non-causal relative to the wave excitation (Clément and Babarit, 2012), i.e. it depends on the future values of the wave excitation. In the frequency domain, it is possible to find the optimal latched mode duration, since by definition the future is the same as the past. Nevertheless, this situation practically applies only to regular waves. Therefore, in the real world, sub-optimal solutions obtained in the time domain are necessary, which rely on Pontryagin's principle (Clément and Babarit, 2012).

Over the years, latching control has been analysed by different researchers due to the associated large increase in energy production over resistive control. Some works have been carried out in the frequency domain (Bjarte-Larsson and Falnes, 2006; Valério *et al.*, 2007; Falcão, 2008). Hoskin and Nichols (1986) have first proposed the application of Pontryagin (1987) principle to find the optimal control of WECs. Since their work, Korde (2002), Babarit *et al.* (2004), Babarit and Clément (2006), Clément and Babarit (2012) and Henriques *et al.* (2016) have applied Pontryagin's principle to find the optimal control action (i.e. latch or delatch) at each time instant using time-domain simulations. This approach requires information of the excitation force over a future time horizon. A solution is found by firstly initiating the control input randomly. Subsequently, the state equations are integrated forward in time and then the costate equations backward in time until convergence is achieved for the control input. Nevertheless, Korde (2002), Babarit *et al.* (2004), Babarit and Clément (2006) and Clément and Babarit (2012) assume knowledge of the excitation force over the whole wave trace they consider. This is unrealistic for a practical application, since the wave elevation can predicted with confidence up to 15 s (Fusco and Ringwood, 2010b) and the computational cost would prevent a real-time implementation. Henriques *et al.* (2016) provide a more practical approach based on a receding horizon with a realistic duration, where data from the previous time step is used to initialize the estimates at the current time step.

The main advantage of latching control is its increase in energy absorption being achieved without any reactive power flows (Clément and Babarit, 2012). Thus, the generator does not need to act as a motor and the system does not need to feed power into the sea for part of the wave cycle. As a result, the cost of the electrical and power-electronics components can be greatly reduced. Conversely, the design of the breaking system does not need to be very complex. Additionally, since the device is latched when its velocity is zero, the magnitude force the breaking system experiences is never excessive. However, maintenance and fatigue life need to be considered in the design of the breaking system. For OWCs, the breaking system is represented by a high-speed valve in series with the turbine (Henriques *et al.*, 2016). Nevertheless, its latching action will not be immediate due to the air chamber stiffness (Henriques *et al.*, 2016).

Budal and Falnes (1977), Bjarte-Larsson and Falnes (2006), Valério *et al.* (2007) and Falcão (2008) have treated a single-degree-of-freedom device in heave, while Korde (2002), Babarit *et al.* (2004), Babarit and Clément (2006), Clément and Babarit (2012) and Henriques *et al.* (2016) have considered the interactions from multiple degrees of freedom on the same WEC. In particular, latching control has been applied to different classes of WECs, including OWCs (Korde, 2002; Henriques *et al.*, 2016), a submerged point absorber (Valério *et al.*, 2007), heaving buoys (Budal and Falnes, 1977; Bjarte-Larsson and Falnes, 2006; Babarit *et al.*, 2004; Babarit and Clément, 2006; Clément and Babarit, 2012), and point absorbers with internal moving masses (Babarit *et al.*, 2004; Babarit and Clément, 2006; Clément and Babarit, 2012). To date, no work has been published on the application of latching control to an array of WECs. In fact, the considerations on optimal phase conditions on which latching control is based do not hold for the treatment of multiple WECS (Falnes, 1980; Thomas and Evans, 1981).

Declutching control is a similar discrete strategy introduced by Salter *et al.* (2002) that consists in the connection and disconnection (or declutching) of the control system during part of the wave cycle in order to control the phase of the response of the WEC. The application of no control force during part of the wave cycle results in motions of greater magnitude and with a phase matching the wave excitation during the remaining part of the wave cycle, which results in an overall increase in energy absorption. In hydraulic PTO systems, the declutching is obtained through the activation of a by-pass valve, while in electromechanical PTO systems it can be achieved through a clutch mechanism. Babarit *et al.* (2009) have proposed the adoption of Pontryagin's principle for the determination of the optimal declutching timing in irregular waves. They have shown that with a hydraulic PTO system using declutching control results in a simpler, cheaper PTO unit and an increase in performance over resistive control. Both latching and declutching control achieve phase control without any associated negative, or reactive, power flows, although a very small amount of energy is required for the operation of the latching and declutching mechanisms, respectively. Clément and Babarit (2012) have shown that combining latching and declutching control can result in a substantial increase (by as much as 2.6 times as compared with latching control and 25 times over declutching control in regular waves) in energy absorption over either method. However, these studies do not account for realistic displacement constraints and a realistic computational time for a real-time implementation.

Alternative discrete control types based on bang-bang approaches have been proposed by Li *et al.* (2012) and Abraham and Kerrigan (2013), with the former relying on dynamic programming and the latter on a non-linear system optimized with a variation of the projected gradient scheme.

One of the most promising approaches to the control of WECs that has been the centre of academic research over the past decade is model predictive control. Model predictive control (Bordons and Camacho, 2007) is a control scheme with a successful record in chemical engineering, plant and process control. It consists in the selection of an optimal control action at every sample time by running a quadratic optimization of the dynamic model over a finite future time horizon with a moving window, thus relying on an element of prediction. The application of model predictive control to WECs has been first proposed by Gieske (2007). The first studies on model predictive control relied on linear state-space models of the WEC dynamics (Brekken, 2011; Hals *et al.*, 2011; Cretel *et al.*, 2011; Richter *et al.*, 2014; Li and Belmont, 2014a). An observer of the fictitious radiation states has been included by Andersen *et al.* (2014). Ferri *et al.* (2014) have included the consideration of structural loadings and fatigue lifetime within the cost function. Non-linear mooring effects have been considered by Richter *et al.* (2013) and Amann *et al.* (2015), with the former approach presenting implementation issues due to the high cost associated with a real-time optimization of a non-convex problem and the second approach relying on an estimator and linear model predictive control instead. Implementations for the decentralised and centralised control of a small array of WECs have also been proposed by Oetinger *et al.* (2014) and Li and Belmont (2014b) and Oetinger *et al.* (2015), respectively.

The main advantage of model predictive control is the simple inclusion of displacement, velocity and force constraints and a reduction of the reactive power flow in its framework. As a result, model predictive control typically shows greater energy absorption than other control strategies when realistic constraints are active (Hals *et al.*, 2011; Cretel *et al.*, 2011; Richter *et al.*, 2014). Nevertheless, no studies to date show the performance of model predictive control in energetic wave conditions, when realistic non-linear effects become important. Under those conditions, the accuracy of the dynamic model used by model predictive control is likely to drop significantly, with negative consequences on system performance (Wave Energy Scotland, 2016). Furthermore, model predictive control is likely to suffer significantly from measurement noise in the wave elevation as well as in its forecast, as shown by Tona *et al.* (2015).

Most of the approaches described above are not adaptive to changes in the system dynamics with time. These may occur either due to slow marine growth effects or sudden non-critical subsystem failures. Adapting the response of the device optimally to these changes can be advantageous for the reduction of operation and maintenance costs. An adaptive control scheme could enable a more flexible maintenance schedule to be developed, allowing a more cost effective solution in the events of non-critical faults thus without incurring the costs associated with emergency maintenance tasks, which are particularly tricky in marine operations due to the requirements of suitable weather windows. Fusco and Ringwood (2014) propose a strategy that tries to address this issue with the development of a robust hierarchical controller for the reduction of the sensitivity to modelling errors and non-linear effects. An alternative approach that deals with non-linear effects is proposed by Valério *et al.* (2008) that relies on neural networks for the system identification of the Archimedes Wave Swing device. The proposed approach presents a 160% increase in energy absorption over resistive control. In this thesis, machine learning strategies, in particular neural networks and reinforcement learning, are further investigated. The objective is their application to existing control schemes, such as resistive and reactive control, for the creation of adaptive, practical algorithms that present a practical implementation.

In the next sections, resistive and reactive control will be analysed in greater detail, as they will be employed as the basis for the development of the innovative control strategies due to their simplicity.

3.2 Resistive control

Here, resistive or passive control is analysed more in detail. As aforementioned, with this strategy, the PTO force is set to be proportional to the velocity at the PTO. For simplicity, in this thesis a PTO system with a single degree of freedom is considered, although it is possible to extend the following methodology to the treatment of multiple degrees of freedom, as for instance done at Pelamis Wave Power Ltd.

Defining the damping coefficient as B_{PTO}, for point absorber subject to heaving motions, this can be expressed as

$$f_{\mathrm{PTO}} = B_{\mathrm{PTO}}\dot{z}. \tag{3.1}$$

The corresponding instantaneous absorbed power is given by (Korde and Ringwood, 2016)

$$P = B_{\mathrm{PTO}} f_{\mathrm{PTO}}, \tag{3.2}$$

where the PTO force may present saturation constraints.

Ignoring the effects associated with PTO force saturation, in the frequency domain using (2.52) it is possible to show that the optimal control damping coefficient for regular waves with circular frequency ω is given by (Falnes, 2005; Korde and Ringwood, 2016)

$$B_{\mathrm{PTO,opt}} = \sqrt{B_{3,3}^2(\omega) + \left[\omega\left(M_{3,3} + A_{3,3}(\omega)\right) - \frac{C_{3,3}}{\omega}\right]^2}, \tag{3.3}$$

which results in the maximum mean absorbed power

$$\bar{P}_{\mathrm{opt}} = \frac{1}{4} \frac{|f_{\mathrm{e},3}|^2}{B_{3,3}(\omega) + \sqrt{B_{3,3}^2(\omega) + \left[\omega\left(M_{3,3} + A_{3,3}(\omega)\right) - \frac{C_{3,3}}{\omega}\right]^2}}. \tag{3.4}$$

Although it is possible to extend these formulae to the treatment of irregular waves using superposition, realistic saturation force constraints cannot be simply represented in the frequency domain. For this reason, an alternative approach based on optimizations using non-linear, time-domain models is usually preferred (Nambiar *et al.*, 2015; Wave Energy Scotland, 2016).

The simulations are usually based on models similar to (2.78), which present a good compromise between accuracy and computational cost. The mean power is usually measured over a time interval of at least 5 minutes in the same sea state, after the dynamic model is fully initialized. At Pelamis Wave Power Ltd., it was common practice to run the simulations for a time of 6 minutes after initialization, which resulted in a feasible computational cost. Nevertheless, waves were generated for a much longer time for the same sea state (say 6-12 hours). The highest observed wave was then included within the wave trace the controller had to be optimized with. This is a conservative approach that is used to design the controller so that it takes into account the worst case scenario in each sea state. Indeed, not only is the cost function of the optimization scheme designed for the maximization of the energy absorption, but it also accounts for the abidance of displacement constraints. Different optimization schemes have been adopted, with the Simplex being preferred by Nambiar *et al.* (2015) and simulated annealing by Pelamis Wave Power Ltd. Both are non-convex optimization schemes (Arora, 2012), which are preferred due to the non-linear nature of the model of the system dynamics.

The optimal damping coefficient is found in each sea state as determined by discrete values of the significant wave height and the energy wave period. The coefficients are then stored in a look-up table, which will be used by the controller on the full-scale WEC to select the optimal parameter for each sea state. Nevertheless, it is clear that this approach is overly simplistic. The assumption of stationary wave conditions and the corresponding lack of need for wave forecasting make for a practical implementation at the expense of performance. Furthermore, the system may be significantly affected by modelling errors and the controller cannot adapt to changes in the system dynamics with time.

3.2.1 Special cases

The case studies analysed in this thesis present particular implementations of resistive control that are treated in the following sections.

3.2.1.1 Two-body point absorber

Although the RM3 point absorber presented in Section 2.3.2 presents two bodies, each with a degree of freedom, the PTO can be reduced to a single degree of freedom. In (2.95), the PTO force acting on each body in opposite directions is given by

$$f_{\mathrm{PTO}} = B_{\mathrm{PTO}} \left(\eta_3 - \eta_9 \right), \tag{3.5}$$

where η_3 and η_9 are the heaving displacements of the float and reaction plate, respectively. Realistic saturation constraints can be applied to the control force.

3.2.1.2 Seabased point absorber

The Seabased point absorber presents a more complex PTO force due to its direct drive PTO system. A simple model that accounts for the overlap between translator and stator has been proposed by Eriksson *et al.* (2007). The electromotive force is proportional to stator current i_{s} and active area A_{fac}

$$f_{\mathrm{PTO}} = k_{\tau} A_{\mathrm{fac}}(y) i_{\mathrm{s}}, \tag{3.6}$$

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 $i_{\mathrm{s}} = b\dot{y}$, with b being a constant, then (3.6) becomes

$$f_{\mathrm{PTO}} = k_{\tau} b A_{\mathrm{fac}}(y) \dot{y}, \text{ or} \tag{3.7a}$$

$$f_{\mathrm{PTO}} = B_{\mathrm{PTO}} A_{\mathrm{fac}}(y) \dot{y}, \tag{3.7b}$$

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

$$A_{\mathrm{fac}}(y) = \begin{cases} 0 & \text{if } |y| \geq 0.5(l_{\mathrm{p}} + l_{\mathrm{s}}), \quad (3.8a) \\ 1 & \text{if } |y| \leq 0.5(l_{\mathrm{p}} - l_{\mathrm{s}}), \quad (3.8b) \\ \left[0.5(l_{\mathrm{p}} + l_{\mathrm{s}}) - |y|\right]/l_{\mathrm{s}} & \text{else}, \quad (3.8c) \end{cases}$$

with l_{p} and l_{s} being introduced in Section 2.3.3.

3.3 Reactive control

Reactive control can be considered to be an extension of resistive control, with the PTO force now having a term proportional to the velocity and one to the velocity at the PTO:

$$f_{\mathrm{PTO}} = B_{\mathrm{PTO}} \dot{z} + C_{\mathrm{PTO}} z \tag{3.9}$$

for a single degree of freedom system, where C_{PTO} is the PTO stiffness coefficient. The stiffness term contributes to the control of the phase of the WEC response. Note that as an alternative means of phase control, it is possible to employ an inertia term, as for instance investigated by Price (2009). However, Hansen *et al.* (2013) has shown that the stiffness term results in a more robust control with a flatter response.

The addition of the stiffness term results in negative power flow during part of the wave cycle. This corresponds to power being fed into the waves in order to change the response of the device with an increase in power absorption in the remaining part of the wave cycle, for an overall maximization of the extracted energy. The negative power is in fact achieved through the generator acting as a motor for part of the wave cycle, which requires a special design. In practice, this has been successfully achieved by a number of companies at both model and full scale (Wave Energy Scotland, 2016).

Using (2.52) in the frequency domain in regular waves with circular frequency ω and ignoring the effects associated with PTO force saturation, the optimal control damping and stiffness coefficients can be expressed for a point absorber limited to heave as (Falnes, 2005; Korde and Ringwood, 2016)

$$B_{\mathrm{PTO,opt}} = B_{3,3}(\omega), \tag{3.10a}$$

$$C_{\mathrm{PTO,opt}} = \omega^2 \left(M_{3,3} + A_{3,3}(\omega)\right), \tag{3.10b}$$

which correspond to the maximum mean absorbed power

$$P = \frac{|f_{\mathrm{e},3}|^2}{8 B_{\mathrm{PTO}}}. \tag{3.11}$$

Therefore, it is clear that for optimal power absorption the controller should match the radiation impedance of the WEC; thus the name impedance-matching.

Even though reactive power results in much greater energy absorption theoretically, as shown in the next section, in practice this this is associated with values of the control force, body displacement, structural loading and negative power flows that are not feasible in practice. For this reason, the PTO force is likely to reach the saturation limit and displacements constraints are likely to be exceeded in all but the mildest sea states. A solution is the increase of the damping coefficient and a reduction of the magnitude of the stiffness coefficient. Furthermore, point absorbers are likely to present a negative optimal control stiffness coefficient (Falcão, 2008; Wave Energy Scotland, 2016). If the magnitude of the PTO coefficient exceeds the actual restoring stiffness of the WEC, the system will become unstable (Wave Energy Scotland, 2016). As a result, simulations in the time domain are necessary for the selection of suitable coefficients with reactive control (Nambiar *et al.*, 2015), particularly if sea states are assumed to be stationary.

3.4 Power calculation

Real PTO systems present energy losses associated with the conversion of energy. In general, the smaller the number of energy conversion stages, the higher the efficiency of the PTO system, with direct drive units promising greater efficiency than electromechanical systems, which in turn are expected to outperform hydraulic systems (Cruz, 2008; Castellini *et al.*, 2014).

In this thesis, a very simplistic model is used to account for the PTO losses, with a unique figure being used for the efficiency of the overall PTO system, η, as done by Nambiar *et al.* (2015). This approach is very computationally efficient and does not affect the methodology developed in this work. However, during later stages of the design process, more detailed models will be necessary.

With this method, the instantaneous generated power is approximated as (Nambiar *et al.*, 2015)

$$P(t) = \begin{cases} \eta f_{\mathrm{PTO}}(t)\dot{z}(t) & \text{if } f_{\mathrm{PTO}}(t)\dot{z}(t) > 0, \quad (3.12a) \\ \frac{1}{\eta} f_{\mathrm{PTO}}(t)\dot{z}(t) & \text{otherwise .} \quad (3.12b) \end{cases}$$

This equation accounts for both positive (i.e. generated) and negative (i.e. supplied to the device) power flows.

3.5 Case study

In this section, state-of-the-art resistive and reactive control with stationary sea states are applied to the single degree point absorber introduced in Section 2.3.1.

3.5.1 Regular waves

First of all, regular waves are analysed with no PTO force saturation and no displacement constraints. As a result, it is possible to obtain the optimal damping and stiffness coefficients in the frequency domain with (3.3) and (3.10a) and (3.10b) for resistive and reactive control, respectively. The corresponding mean absorbed power in regular waves with unit amplitude is computed with (3.2) and (3.11), respectively. From the mean power, it is possible to compute the capture width, which indicates the width of the incoming wave front with a power corresponding to the one absorbed by the device (Cruz, 2008). As a result, the capture width can be obtained as

$$\mathcal{L}\left(\omega,\beta\right)=\frac{P}{P_{\mathrm{w}}}. \tag{3.13}$$

For an axisymmetric device, as those treated in this thesis, the dependence on the wave direction is dropped. In (3.13), P_{w} is the mean power per unit crest of the incoming waves (Cruz, 2008), which is given by

$$P_{\mathrm{w}}=\frac{1}{2}\rho g a^{2}c_{\mathrm{g}}, \tag{3.14}$$

where a is the wave amplitude (1 m here) and c_{g} the wave group velocity, which is obtained as follows, assuming deep water:

$$c_{\mathrm{g}}=\frac{1}{2}\frac{g}{\omega}. \tag{3.15}$$

The capture width ratio expresses the ratio of the capture width and a typical device dimension (Cruz, 2008). In the case of an axisymmetric point absorber, the diameter (10 m in this case) is typically used as parameter, with the capture width ratio being by $\mathcal{L}/D$.

The variation with non-dimensional wave frequency of the capture width ratio of analysed point absorber when resistive and reactive control are applied can be seen in Figure 3.1. The axes are limited to produce a clearer plot. The capture width ratio associated with reactive control and no constraints tends to infinity as the wave frequency tends to zero.

Subsequently, the influence of PTO force saturation and displacement constraint is investigated in the time-domain. A single sea state with regular waves of unit amplitude

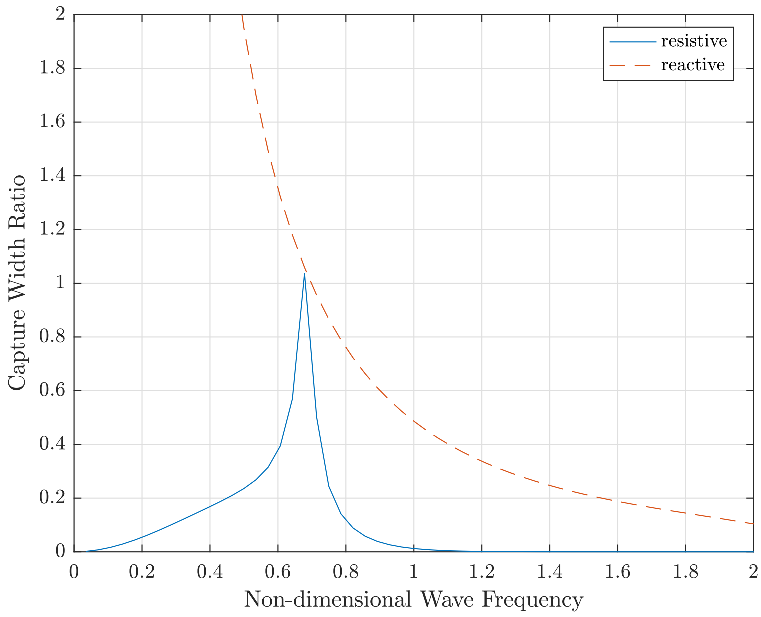


Figure 3.1: Variation in the capture width ratio of the analysed point absorber with non-dimensional wave frequency. The curves for both resistive and reactive control are applied.

and a period of 8 s is considered. In this wave trace, the response of the device is analysed when resistive and reactive control are applied with no constraints, with a force saturation limit of 200 kN and unbounded displacement, and with no force saturation but with the displacement magnitude constrained to 1 m. A Simplex optimization algorithm (Arora, 2012) is used for the determination of the optimal PTO damping coefficient and PTO damping and stiffness coefficients for resistive and reactive control, respectively. The PTO damping coefficient was bounded to a maximum value of 10 MNs/m to prevent numerical errors. The computed values can be seen in Table 3.1.

The results of this analysis are displayed in Figures 3.2 and 3.3 for resistive and reactive

Table 3.1: Optimal controller damping and stiffness coefficients for all analysed control types and limits on the PTO force and body displacement in regular waves of unit amplitude and a period of 8 s.

control type	max $\|f_{\mathrm{PTO}}\|$ (kN)	max $\|z\|$ (m)	$B_{\mathrm{PTO,opt}}$ kNs/m	$C_{\mathrm{PTO,opt}}$ kN/m
resistive	-	-	305.536	-
resistive	200	-	3232.143	-
resistive	-	1	464.639	-
reactive	-	-	100.652	-226.505
reactive	200	-	6781.990	-2615.230
reactive	-	1	404.194	-119.478

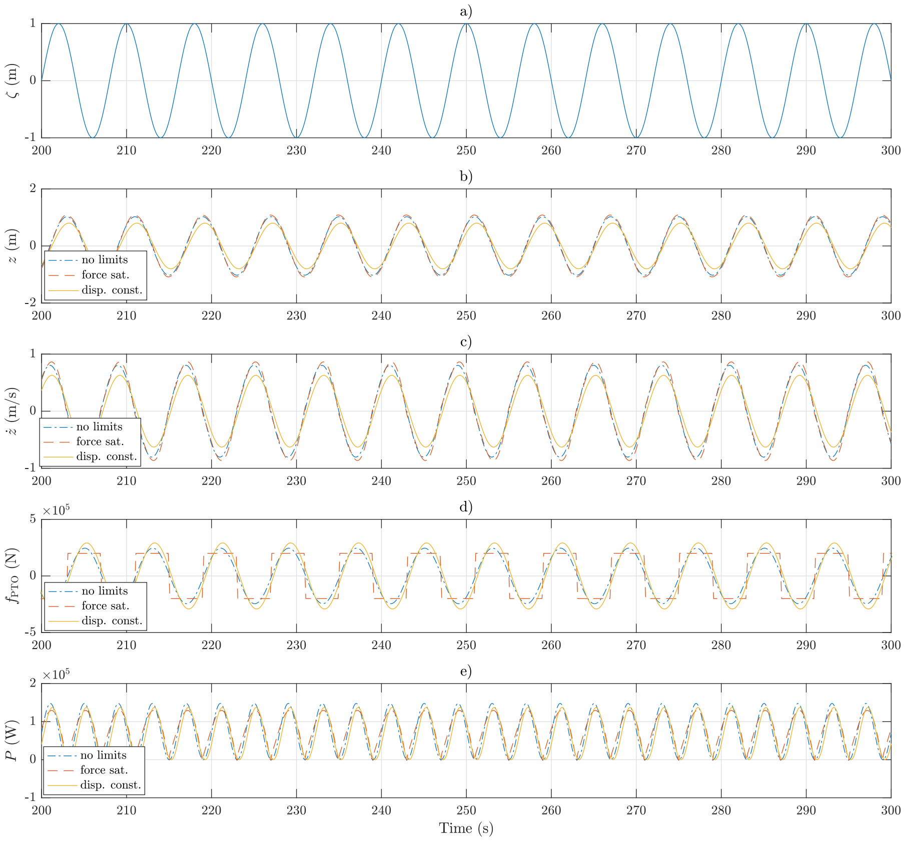


Figure 3.2: Wave elevation (a), body vertical displacement (b) and velocity (c), PTO force (d) and corresponding absorbed power (e) when the WEC is passively controlled in regular waves of unit amplitude and a period of 8 s. The PTO damping coefficient and the force and displacement limits shown in Table 3.1 are used.

control, respectively. The figures show the body displacement and velocity, the PTO force and the absorbed power after the system is fully initialized. No curves are shown for the filtered, mean power. In these figures, it is possible to distinguish the curves corresponding to the case of no constraints, PTO force saturation and displacement limit.

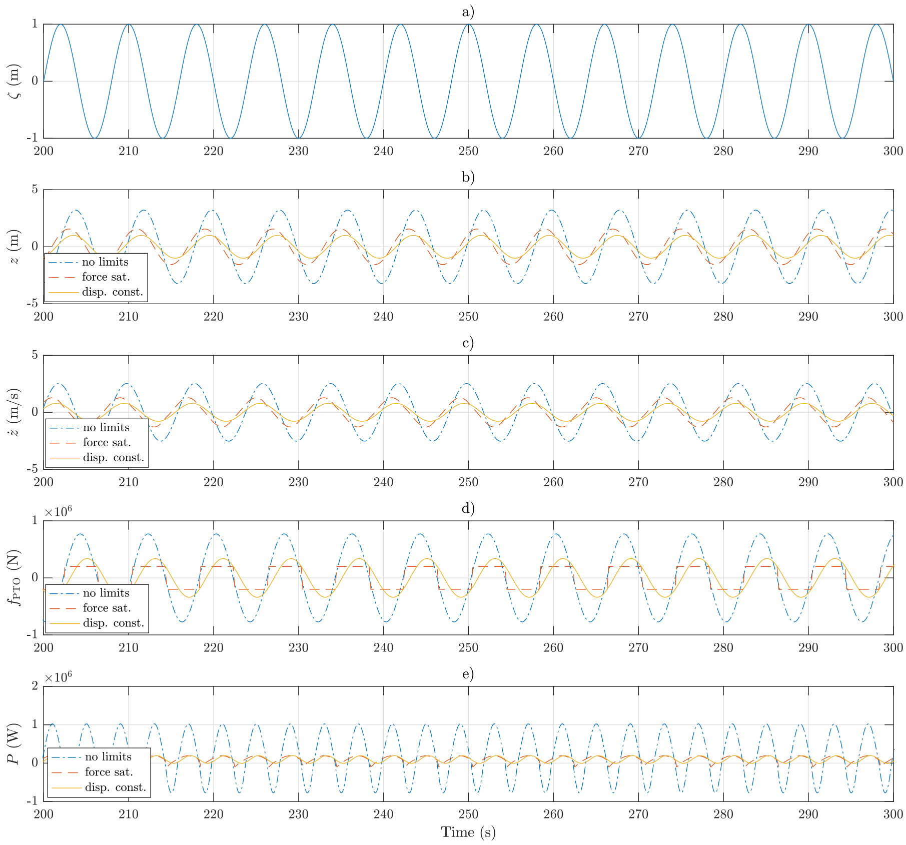


Figure 3.3: Wave elevation (a), body vertical displacement (b) and velocity (c), PTO force (d) and corresponding absorbed power (e) when the WEC is actively controlled in regular waves of unit amplitude and a period of 8 s. The PTO damping and stiffness coefficients and the force and displacement limits shown in Table 3.1 are used.

Table 3.2: Optimal controller damping and stiffness coefficients for all analysed control types and limits on the PTO force and body displacement in irregular waves with a JONSWAP spectrum, $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 8$ s.

control type	$\max\lvert f_{\mathrm{PTO}}\rvert$ (kN)	$\max\lvert z\rvert$ (m)	$B_{\mathrm{PTO,opt}}$ kNs/m	$C_{\mathrm{PTO,opt}}$ kN/m
resistive	-	-	408.118	-
resistive	200	-	473.042	-
resistive	-	1	657.027	-
reactive	-	-	187.036	-320.794
reactive	200	-	440.246	-274.394
reactive	-	1	906.016	-415.915

3.5.2 Irregular waves

Similarly to regular waves, an analysis has been conducted in a 500-s-long wave trace in irregular waves with a JONSWAP spectrum with a significant wave height of 2 m and a peak wave period of 9.25 s, corresponding to an energy wave period of 8 s from spectral analysis (Holthuijsen, 2007). As before, the response of the device is analysed when resistive and reactive control are applied with no constraints, with a force saturation limit of 200 kN and unbounded displacement, and with no force saturation but with the displacement magnitude constrained to 1 m. The obtained optimal PTO damping and stiffness coefficients can be seen in Table 3.2.

The results of this analysis are displayed in Figures 3.4 and 3.5 for resistive and reactive control, respectively. The figures show the body displacement and velocity, the PTO force and the absorbed power after the system is fully initialized (starting from a time of 200 s). No curves are shown for the filtered, mean power. In these figures, it is possible to distinguish the curves corresponding to the case of no constraints, PTO force saturation and displacement limit.

3.5.3 Discussion

As is clear from Figure 3.1, reactive control has the potential of much higher energy absorption over resistive control. From Table 3.1, it is interesting to notice that the optimal PTO stiffness coefficient presents a negative value, as predicted by Falcão (2008) and Wave Energy Scotland (2016) for point absorbers.

However, as shown by Figures 3.2 and 3.3, reactive control also results in significantly greater body displacements and PTO force as well as substantial negative power flows. In order to meet the body displacement constraints, the controller increases the PTO damping coefficient for both resistive and reactive control, and decreases the magnitude of the PTO stiffness coefficient for reactive control, as can be seen in Table 3.1. The controller behaviour in regular waves under PTO saturation is more interesting, with

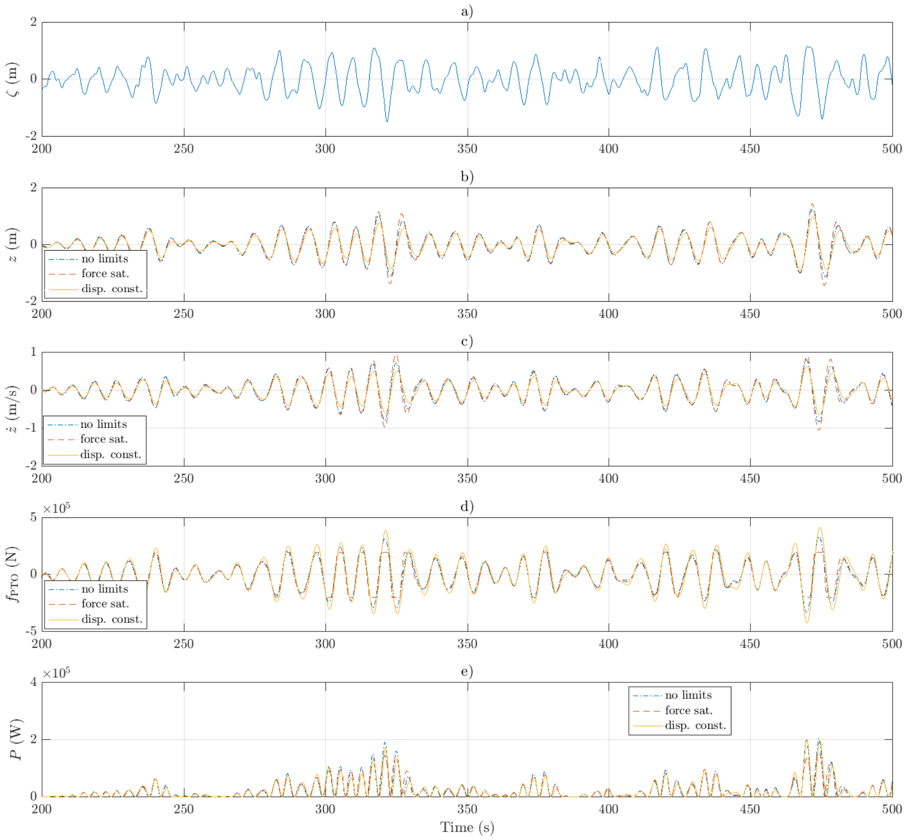


Figure 3.4: Wave elevation (a), body vertical displacement (b) and velocity (c), PTO force (d) and corresponding absorbed power (e) when the WEC is passively controlled in irregular waves with a JONSWAP spectrum, $H_s = 2$ m and $T_e = 8$ s. The PTO damping coefficient and the force and displacement limits shown in Table 3.1 are used.

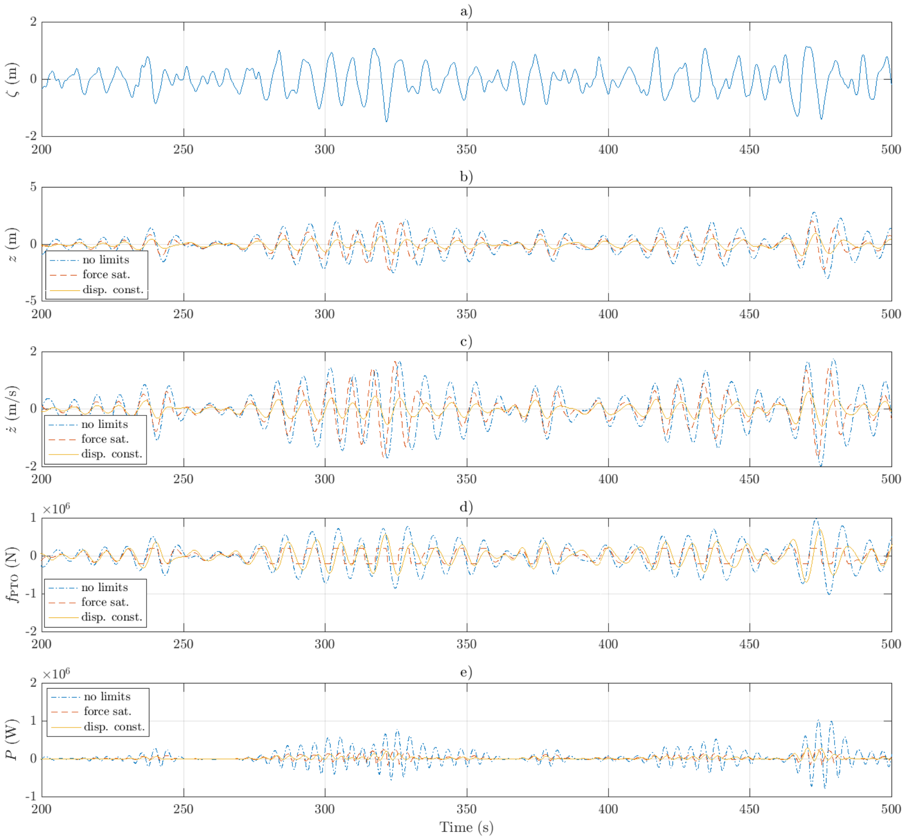


Figure 3.5: Wave elevation (a), body vertical displacement (b) and velocity (c), PTO force (d) and corresponding absorbed power (e) when the WEC is actively controlled in irregular waves with a JONSWAP spectrum, $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 8$ s. The PTO damping and stiffness coefficients and the force and displacement limits shown in Table 3.1 are used.

the control force converging towards a bang-bang type of control. This is clear from the shape of the PTO force in Figures 3.2 and 3.3, which resembles a square wave.

In irregular waves (Figures 3.4 and 3.5), the bang-bang behaviour is no-longer observed when the force saturates, since the limits apply to a much shorter instance related to particularly energetic waves. Hence, it is more advantageous for the controller to maximize energy absorption over the whole wave trace, whilst meeting the constraint only for the highest wave. This results in a small change in PTO coefficients and thus a small drop in performance. Conversely, the performance of the controller is much more affected by limits on the body displacement than in regular waves, as the coefficients are tailored to meet the strictest constraints associated with the highest wave in the analysed wave trace for the optimization. This behaviour was observed in previous studies (Cretel *et al.*, 2011; Richter *et al.*, 2014) and is the main reason for the research in real-time control schemes, such as model predictive control, which can better deal with constraints due to their optimization of the response on a wave by wave basis. Finally, from Figures 3.4 and 3.5, it is also interesting to notice the transport of wave power in packets, which are known as wave groups.

3.6 Chapter summary

In this chapter, state-of-the-art technologies for the control of WECs are reviewed. A distinction is made between strategies that have been successfully implemented on WEC prototypes and schemes that show promise of superior performance, but are still the subject of academic studies. Of the latter strategies, model predictive control, a real-time technique, is considered to have great potential. Nevertheless, due to the industrial nature of this project, the former group of schemes is selected for the development of innovative algorithms. These methods are based on the assumption of stationary wave conditions (known as sea states) for periods of 15 to 30 minutes, which relies on wave data statistics. The PTO force is then modelled as a damping or the combination of damping and stiffness terms for passive and active (also known as resistive and reactive) control, respectively. For each sea state, the optimal coefficients are found using simulations. The cost function is based on the maximization of energy absorption while considering realistic constraints on the PTO force and body displacement. A case study is presented using the model of point absorber constrained to heave, which was introduced in the previous chapter. The superior performance associated with reactive control is shown, although the limitations in performance of these strategies when displacement constraints are active is discussed.

From the literature review, it is clear that the existing control strategies for WECs can be adaptive to changes in wave conditions, but not changes in their dynamics.

Nevertheless, marine growth seriously modifies the response of the WECs during their lifetime. Additionally, it would be advantageous for the WEC to adapt to non-critical subsystem failures so that the maintenance schedule can be optimized. For these reasons, machine learning strategies will be investigated for the development of model-free control schemes for WECs. In the next chapter, neural networks, a class of supervised learning schemes, are treated, while reinforcement learning, a framework to make decisions belonging to the unsupervised learning class, is discussed in Chapter 4.

Chapter 4

Reinforcement learning

4.1 Background

Reinforcement learning is a class of nature-inspired algorithms, which have become very popular within the robotics and machine learning communities (Mnih *et al.*, 2015). The technique belongs to the class of unsupervised learning strategies and it is based on the idea of learning from experience coupled with the principle of reward and punishment for survival and growth (Khan *et al.*, 2012). The theory of reinforcement learning is treated in detail in Sutton and Barto (1998), which is the main introductory book on the subject, and Busoniu *et al.* (2010). A review of modern applications can be found in Khan *et al.* (2012) and Littman (2015). Furthermore, modern approaches, including function approximation and newer algorithms, are addressed in detail in Geramifard *et al.* (2013).

In reinforcement learning (Sutton and Barto, 1998), an agent, which is in a particular *state* s, interacts with the surrounding environment by taking an *action* a. The agent then moves to a new state, s', and the action is followed by a *reward*, r, depending on its outcome. The action selection process is modelled as a Markov decision process based on the *value function*, which expresses the estimate of the future reward. The agent is expected to learn an optimal behaviour, known as *policy*, over time for the maximization of the total reward. This process is shown graphically in Figure 4.1, which is taken from Sutton and Barto (1998). In control terminology, Figure 4.1 can be explained as follows (Khan *et al.*, 2012):

- the state signal describes the state of the environment and agent;
- the action signal represents the control input;
- the reward signal is a feedback signal.

If the agent selects an action based purely on the aim of maximising the reward function (i.e. exploiting the environment), it will never visit states other than the usual ones, and these other states may in fact result in higher rewards. This is known as the issue of *exploration* versus *exploitation*. Hence, it is still beneficial to adopt an approach that ensures some exploration at the expense of exploitation, particularly for the initial

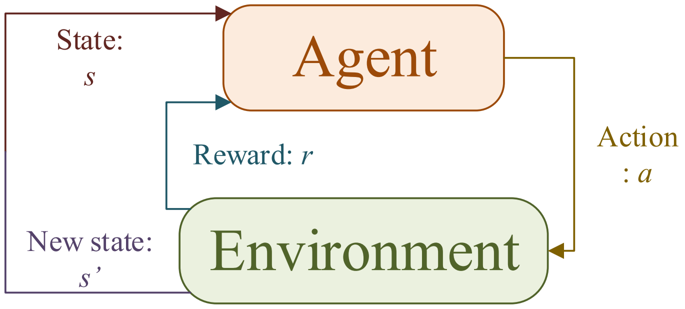


Figure 4.1: Diagram of the reinforcement learning work flow (Sutton and Barto, 1998).

stages. Once the simulation has been initialized, the balance may be shifted towards exploitation. Exploration strategies will be treated in Section 4.3.

Before moving to the treatment of different schemes, some of the reinforcement learning terms that have been introduced are explained thoroughly, since they will be employed throughout this and future chapters.

- The policy, usually represented as π, is the behaviour of the agent at a particular time. It may be stochastic or deterministic. A greedy policy, i.e. such that it maximizes the value function, is typically specified, with a strong link to the exploration strategy. The reinforcement learning process will lead to an optimal policy with time (Sutton and Barto, 1998).
- The reward function can be considered as an inverse cost function. It is defined on the basis of the goal the agent is expected to achieve. A discount factor can be used to give more importance to either immediate or more future-oriented reward. Designing an appropriate reward function is particularly challenging, since it can be very difficult to determine what actions should be rewarded in complex problems. For this reason, *apprenticeship learning* has been developed (Abbeel, 2008), where the reward function is derived from a statistical study of the actions taken by an expert while performing the task.
- Two types of value functions are used in the reinforcement learning literature: the state value function, $V(s)$, and the state-action value function, $Q(s,a)$. The former is preferred when a model of the environment is available, e.g. in dynamic programming; the latter is used when the model of the environment is not known, e.g. in the Monte-Carlo and temporal difference methods (Sutton and Barto, 1998). The value function represents the prediction of the future reward for a given state or state-action pair. Conversely, the reward function returns the immediate reward. Therefore, decisions on the action selection are based on the value function, rather than the reward function, since it provides an estimate of

the total future reward expected after landing in a specific state.

Reinforcement learning methods can be divided into three main categories: *dynamic programming*, *temporal difference* and *Monte-Carlo methods* (Sutton and Barto, 1998). Dynamic programming is mathematically well developed, but needs a full model of the environment in order to determine a suitable policy (Sutton and Barto, 1998; Busoniu *et al.*, 2010). Thus, this strategy is not treated in this work because the aim is the development of a model-free controller. Conversely, with Monte-Carlo and temporal difference methods learning can occur from direct observations of the environment. While Monte-Carlo techniques need to wait for the end of the task before updating the value function, temporal difference schemes can learn on-line like dynamic programming (Sutton and Barto, 1998; Busoniu *et al.*, 2010). Monte-Carlo methods are treated in Section 4.4, while temporal difference strategies in Section 4.5. First of all, Markov decision processes are described in Section 4.2, introducing the Bellman equation, while exploration strategies are considered in Section 4.3.

4.2 Markov decision processes

The theory of Markov decision processes is taken from Sutton and Barto (1998), Lagoudakis and Parr (2003), Busoniu *et al.* (2010) and Geramifard *et al.* (2013) so that these references will no longer be repeated in this section. A Markov decision process is defined as a tuple of the form $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \gamma)$ where $\mathcal{S} = \{s_1, s_2, \ldots, s_I\}$ is a finite set of I states and $\mathcal{A} = \{a_1, a_2, \ldots, a_J\}$ a finite set of J actions. $\mathcal{P}$ is a Markov transition model, with $\mathcal{P}(s, a, s') = p(s \xrightarrow{a} s')$ being the probability of transitioning to state s' when taking action a in state s. $\gamma \in [0, 1]$ is the discount factor for future rewards. It is assumed that the Markov decision process has an infinite horizon so that future rewards are discounted exponentially. $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathrm{I\!R}$ is the reward function, where $R(s, a, s')$ is the reward corresponding to the transition $s \xrightarrow{a} s'$. The notation can be simplified to $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathrm{I\!R}$, with the predicted reward for state-action pair (s, a) expressed as

$$\mathcal{R}(s, a) = \sum_{s' \in \mathcal{S}} \mathcal{P}(s, a, s') R(s, a, s'). \tag{4.1}$$

For a Markov decision process, a stationary policy π is a mapping between states and actions: $\pi : \mathcal{S} \mapsto \Omega(\mathcal{A})$, where $\Omega(\mathcal{A})$ is the set of all probability distributions over $\mathcal{A}$. Hence, $\pi(a; s)$ indicates the probability that policy π selects action a in state s. A particular case is represented by a stationary deterministic policy that results in a single action choice per state. In this case, the mapping reduces to $\pi : \mathcal{S} \mapsto \mathcal{A}$ from states to actions, so that $\pi(s)$ indicates the action taken in state s.

The state-action value function $Q^{\pi}(s,a)$ of any policy π is defined over all possible combinations of states and actions. As aforementioned, the value function indicates the expected, discounted, total reward that will follow the selection of action a in state s and following policy π thereafter:

$$Q^{\pi}(s,a) = E_{a_t \sim \pi; s_t \sim \mathcal{P}} \left(\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a \right), \tag{4.2}$$

where E is the expected reward, r_t the reward at time t and s_0 and a_0 as the starting state and action, respectively. For all state-action pairs, the exact state-action values can be found by solving the linear system of the Bellman equations (Bellman, 1957):

$$Q^{\pi}(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s,a,s') \sum_{a' \in \mathcal{A}} \pi(a';s') Q^{\pi}(s',a'). \tag{4.3}$$

This system can be expressed in matrix form as

$$Q^{\pi} = \mathcal{R} + \gamma \boldsymbol{P} \boldsymbol{\Pi}_{\pi} Q^{\pi}, \tag{4.4}$$

where Q^{π} and $\mathcal{R}$ are vectors of size $(|\mathcal{S}||\mathcal{A}|, 1)$. $\boldsymbol{P}$ is a stochastic matrix of size $(|\mathcal{S}||\mathcal{A}|, |\mathcal{S}|)$ that includes the transition model of the process

$$\boldsymbol{P}\left((s,a),s'\right) = \mathcal{P}(s,a,s'). \tag{4.5}$$

$\boldsymbol{\Pi}_{\pi}$ is a stochastic matrix of size $(|\mathcal{S}|, |\mathcal{S}||\mathcal{A}|)$ that describes the policy

$$\boldsymbol{\Pi}_{\pi}\left(s', (s',a')\right) = \pi(a';s'). \tag{4.6}$$

The resulting linear system

$$\left(\boldsymbol{I} - \gamma \boldsymbol{P} \boldsymbol{\Pi}_{\pi}\right) Q^{\pi} = \mathcal{R} \tag{4.7}$$

can be solved analytically or iteratively in order to obtain the exact state-action values, where $\boldsymbol{I}$ is the identity matrix in this case.

For every Markov decision process, there exists an optimal deterministic policy, π^*, which maximizes the expected, total, discounted reward from any initial state. This corresponds to

$$\pi^* = \arg\max_{\pi} Q^{\pi}(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{4.8}$$

Hence, the search for the optimal policy can be restricted only to the space of deterministic policies.

In Monte-Carlo and temporal difference methods, the transition matrix $\boldsymbol{P}$ is not known if a model of the system is not employed. Hence, the state-action value function is updated using information derived from direct observations of the environment. In particular, with on-line temporal difference schemes, the update occurs at every step of the algorithm, while with batch-mode algorithms the Q function is updated using a number of samples stored in a batch. The difference between on-line and batch-mode temporal difference strategies is treated in Section 4.5. With Monte-Carlo methods, the state-action value function is updated at the end of each episode.

4.3 Exploration strategy

Different strategies have been proposed in order to ensure sufficient exploration at the start of a reinforcement learning problem, whilst shifting the focus to exploitation of the best actions as learning progresses. The most famous techniques are known as ϵ-greedy, Boltzmann and counter-based exploration (Busoniu *et al.*, 2010). Of these strategies, ϵ-greedy and Boltzmann exploration have been investigated, but only the former has been applied to the control of wave energy converters.

Given the current state s, an ϵ-greedy policy selects the action at the start of each step of the algorithm (Busoniu *et al.*, 2010)

$$a = \begin{cases} \arg\max\limits_{a' \in \mathcal{A}(s)} Q(s, a') & \text{with probability } 1 - \epsilon, \quad (4.9a) \\ \text{random action} & \text{with probability } \epsilon, \quad (4.9b) \end{cases}$$

where ϵ is the exploration rate. This means that with probability $1-\epsilon$ the greedy action is selected, i.e. the exploitative action that maximizes the value function and thus the expected total future reward; otherwise an exploitative action is chosen instead. It is important to notice that in (4.9), the action space is described as a function of the state. This is because particular states, e.g. those lying on the boundary of the state space, may present a limited set of actions so as to prevent exceeding the constraints. For instance, this is the case for grid-world problems, with an example being analysed in Section 4.6.1.

In order to ensure greater exploration at the start of reinforcement learning control, whilst focusing on exploitative actions as learning progresses, in this work the exploration rate is decreased with time as follows:

$$\epsilon = \begin{cases} \epsilon_0 & \text{if } N \leq N_\epsilon, \quad (4.10a) \\ \epsilon_0 / \sqrt{N - N_\epsilon} & \text{otherwise.} \quad (4.10b) \end{cases}$$

In (4.10), ϵ_0 is the initial exploration rate, while N_ϵ indicates the minimum number of visits to a specific state before reducing the exploration rate. The number of visits to all *discrete* state-action pairs is stored in the matrix $\boldsymbol{N}$ with size $(|\mathcal{S}|, |\mathcal{A}|)$, with $\mathcal{S}$

indicating the state space and $\mathcal{A}$ the action space. The entry corresponding to the state-action pair (s, a) is given by $\boldsymbol{N}(s, a)$. In (4.10), $N = \sum_{a \in \mathcal{A}} \boldsymbol{N}(s, a)$, i.e. N corresponds to the sum of the number of visits to all actions for the current state. Discrete states are used for the determination of the exploration rate even when function approximation is employed. Note that the decay of the exploration rate presented in (4.10) has been optimized for the application of reinforcement learning to the control of WECs analysed in this work. For more conventional reinforcement learning problems with a greater number of episodes, a slower decay rate may be beneficial. Usually, in these problems the exploration rate decay is dependent on the number of episodes instead (Geramifard *et al.*, 2013).

4.4 Monte-Carlo methods

With Monte-Carlo methods, the reinforcement learning problem is solved based on averaging sample rewards (Sutton and Barto, 1998). In fact, in the terminology of Monte-Carlo techniques, the reward is referred to as *return*. These schemes are defined only for episodic tasks so as to ensure well-defined, complete returns. This means that the experience that the controller observes from interactions with the environment is divided into discrete episodes, which present a defined end no matter what actions are taken. Therefore, the policy is updated only at the end of each episode rather than online. This characteristic makes the application of Monte-Carlo strategies to real-time control problems challenging.

In Monte-Carlo methods, the state-value function for a given policy is estimated by simply averaging over the returns experienced after observing that particular state (Sutton and Barto, 1998). In particular, two different techniques have been developed since the 1940s that differ in their treatment of the visits to a particular state within an episode. In the every-visit Monte-Carlo approach, the value function $V^{\pi}(s)$ is approximated by the mean of the returns following all visits of state s in a set of episodes while following policy π. In the first-visit Monte-Carlo method, the value function is determined from the average over only the first visits to the state. Both techniques converge towards the state-value function $V^{\pi}(s)$ for an infinite number of visits to the state s by the law of large numbers (Sutton and Barto, 1998).

When a model of the environment is not available, Monte-Carlo methods find the optimal policy by alternating a stage of policy evaluation and a stage of policy improvement at the end of each episode. This is shown graphically in Figure 4.2, which is taken from Sutton and Barto (1998). During the policy evaluation phase, the state-action value function is evaluated using the returns stored during the episode. During the policy improvement stage, the policy is updated to the greedy policy with respect to the newly

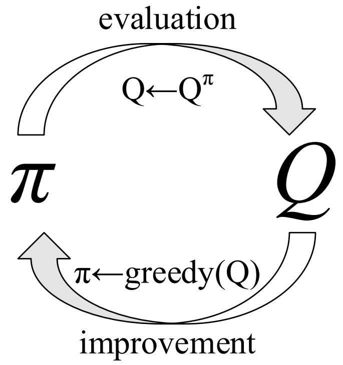


Figure 4.2: Policy iteration in Monte-Carlo methods (Sutton and Barto, 1998).

computed value function, i.e. the policy that maximizes the value function and thus the expected total return. Policy iteration is described in greater detail in Section 4.5.7.1.

Furthermore, Monte-Carlo methods are divided into on- and off-policy schemes. *On-policy* schemes evaluate and improve the policy that is used to make decisions. Conversely, with *off-policy* techniques, the episodes can be generated following a different policy from the one being evaluated. On- and off-policy schemes will be treated in greater detail in Sections 4.5.1 and 4.5.2, respectively, for temporal difference methods. Here, an on-policy, first-visit Monte-Carlo method is treated. In order to ensure all actions are selected infinitely often, *soft*, i.e. meaning that $\pi(s,a) > 0 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, ϵ-greedy strategies have been proposed (Sutton and Barto, 1998). These schemes, like the exploration strategy described in the previous section, gradually shift the policy to a deterministic optimal policy. In Sutton and Barto (1998), all non-greedy actions are selected with probability $\epsilon/|\mathcal{A}(s)|$, while the greedy action with probability $1-\epsilon+\epsilon/|\mathcal{A}(s)|$, where ϵ is the exploration rate. This is similar to the ϵ-greedy exploration strategy described in (4.9). The exploration rate is still calculated according to (4.10), although it is no longer updated at each time step, but rather at the end of each episode.

The work flow can be seen in Algorithm 1, which is adapted from Sutton and Barto (1998). It should be noted that this algorithm assumes discrete states and actions. In Algorithm 1, the first for-loop corresponds to the policy evaluation stage, while the second one to the policy improvement step. $\boldsymbol{R}$ is a list storing all returns, while R indicates the return following the first occurrence of (s,a).

Monte-Carlo methods have been the focus of the early research in reinforcement learning. Famous applications include the determination of the best actions in the multi-arm bandits and black-jack games (Sutton and Barto, 1998). Other problems in the literature include the soap bubble and racetrack examples (Sutton and Barto, 1998). Nevertheless, since the 1990s, reinforcement learning research has focused mainly on

Algorithm 1: On-policy, first-visit Monte Carlo algorithm with ϵ-greedy exploration, adapted from Sutton and Barto (1998).

```
Input: $\epsilon_0$, $N_\epsilon$, $\mathcal{S}$, $\mathcal{A}$
Output: $\pi$
initialize $Q(s,a)$ arbitrarily $\forall s \in \mathcal{S}, a \in \mathcal{A}(s)$;
initialize $\boldsymbol{R}(s,a) \leftarrow [] \ \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$;
initialize $\pi$ with an arbitrary $\epsilon$-soft policy;
while end time not reached do
    run episode following $\pi$;
    for each pair $(s,a)$ in the episode do
        $R \leftarrow$ return following the first occurrence of $(s,a)$;
        Append $R$ to $\boldsymbol{R}(s,a)$: $\boldsymbol{R}(s,a) \leftarrow [\boldsymbol{R}(s,a), R]$;
        $Q(s,a) \leftarrow \mathrm{mean}(\boldsymbol{R}(s,a))$;
    end
    for each $s$ in the episode do
        $a^* \leftarrow \arg\max_{a' \in \mathcal{A}} Q(s,a')$ greedy action;
        for all $a \in \mathcal{A}(s)$ do
            if $a = a^*$ then
                $\pi(s,a) \leftarrow 1 - \epsilon + \epsilon/|\mathcal{A}(s)|$;
            else
                $\pi(s,a) \leftarrow \epsilon/|\mathcal{A}(s)|$;
            end
        end
    end
end
```

temporal difference and dynamic programming methods.

In the context of wave energy control, Monte-Carlo methods have been applied whenever temporal difference methods failed to converge. Their averaging nature ensures Monte-Carlo methods are more robust and are thus able to find the optimal policy even in the most challenging scenarios. In this thesis, Monte-Carlo methods have been applied to the declutching control of a WEC, as can be seen in Chapter 6.

4.5 Temporal difference methods

Temporal difference methods merge the positive aspects of dynamic programming and Monte-Carlo methods (Sutton and Barto, 1998). Like Monte-Carlo methods, temporal difference schemes are independent of a model of the environment dynamics, thus learning from direct observations. Similarly to dynamic programming, these strategies are not episodic, i.e. they can learn on-line without the need to wait for the completion of the task. In particular, temporal difference schemes update their estimate of the value function at the end of each step. Furthermore, temporal difference methods are guaranteed to converge to the actual value function for a fixed policy for a sufficiently small step-size parameter (Sutton and Barto, 1998).

Temporal difference methods have been widely adopted by the robotics and computer science industries. A thorough review of known applications and studies can be found in Khan *et al.* (2012). Littman (2015) focuses on possible future applications in the field of neuroscience. The most famous recent application of reinforcement learning, and temporal difference methods in particular, is the deep reinforcement learning algorithm that Google DeepMind has successfully used to beat the human champion in the game of Go (Mnih *et al.*, 2015).

Temporal difference methods are further subdivided into *on-line* and batch-mode algorithms. With the former strategies, the state-action value is updated at each step (Geramifard *et al.*, 2013). In this work, the popular Sarsa (Rummery and Niranjan, 1994) and Q-learning (Watkins, 1989; Watkins and Dayan, 1992) on-line algorithms are described. Conversely, batch-mode schemes update the state-action value function *off-line* using information from a number, or *batch*, of stored (and previously generated) samples in the form (s, a, r, s') (Geramifard *et al.*, 2013). Here, Neural Fitted Q-iteration (NFQ) (Riedmiller, 2005) and Least-Squares Policy Iteration (LSPI) (Lagoudakis and Parr, 2003) are explained.

4.5.1 Sarsa

Sarsa, which stands for State-Action-Reward-State-Action is an on-line, *on-policy* reinforcement learning algorithm originally proposed by Rummery and Niranjan (1994). This means that the policy used to gather observations is the same as the one used for learning (Geramifard *et al.*, 2013). At each step of the algorithm, the Q-value for the current state-action pair is updated using the resulting immediate reward r and the Q-value for the new state and new action, which will be experienced in the next step (Sutton and Barto, 1998):

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma Q(s',a') - Q(s,a) \right], \tag{4.11}$$

where s' and a' represent the state and action in the next step, respectively. Equation (4.11) can be considered as a numerical approach for the solution of the Bellman equations in (4.4). It is particularly important to visualize the state-action value function Q as a measure of the expected total reward for a particular state and action pair. Note that in this implementation, discrete states and actions are considered. As a result, the state-action value function (or Q-function) can be described by a table. The Sarsa algorithm with linear function approximation for the state space can be found in Section 4.5.4.

In 4.11, α represents the learning rate and γ the discount factor. The discount factor is used to discount future rewards. The learning rate determines the proportion of new and old knowledge that is retained during learning and is calculated here as

$$\alpha = \begin{cases} \alpha_0 & \text{if } \boldsymbol{N}(s,a) \leq N_\alpha, \qquad (4.12\text{a}) \\ \alpha_0/(\boldsymbol{N}(s,a) - N_\alpha) & \text{otherwise}, \qquad (4.12\text{b}) \end{cases}$$

where α_0 and N_α are specified parameters. Equation (4.12) ensures sufficient learning when each state-action pair is visited for the first few times. As learning progresses, older knowledge is given greater importance to limit the impact of sensor noise. Note that the decay of the learning rate presented in (4.12) has been optimized for the application of reinforcement learning to the control of WECs analysed in this work. For more conventional reinforcement learning problems with a greater number of episodes, a slower decay rate may be beneficial. Sarsa is guaranteed to converge for discrete actions and states, a bounded reward variance, the use of a discount factor and a properly decaying learning rate (Singh *et al.*, 2000). The Sarsa algorithm for discrete states is represented in Algorithm 2.

The learning time of the Sarsa algorithm can be greatly reduced with the use of eligibility traces (Sutton and Barto, 1998). Eligibility traces aid the learning algorithm determine the sequence of actions that maximizes the total reward. From a theoretical perspective, they provide a bridge between Monte-Carlo and temporal difference

Algorithm 2: Sarsa: on-line, on-policy reinforcement learning algorithm with discrete, exact states, adapted from Sutton and Barto (1998).

```
Input: $\mathcal{S}$, $\mathcal{A}$, $\alpha_0$, $N_\alpha$, $\gamma$, $\epsilon_0$, $N_\epsilon$
Output: $\pi$
initialize $Q(s,a)$ arbitrarily;
for each episode do
    initialize $\boldsymbol{N} \leftarrow \boldsymbol{0}$;
    initialize $s$;
    get $\epsilon$ with (4.10);
    choose $a$ given $s$ using an $\epsilon$-greedy policy with (4.9);
    for each step in the episode do
        take action $a$, observe $r$, $s'$;
        update $\boldsymbol{N}(s,a) \leftarrow \boldsymbol{N}(s,a) + 1$;
        get $\epsilon$ with (4.10);
        choose $a'$ given $s'$ using an $\epsilon$-greedy policy (4.9);
        get $\alpha$ with (4.12);
        update $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma Q(s',a') - Q(s,a)\right]$ ;
        update $s \leftarrow s'$ ;
        update $a \leftarrow a'$;
    end
end
```

methods (Sutton and Barto, 1998). Although eligibility traces are a very powerful tool, they are not employed in this work because of the nature of the analysed reinforcement learning application, which presents a single episode.

In the context of WEC control, Sarsa has been applied to the resistive control of a point absorber for a comparison with Q-learning and least-squares policy iteration, which have been preferred throughout this work.

4.5.2 Q-learning

The Q-learning algorithm, originally proposed by Watkins (1989) and Watkins and Dayan (1992), is one of the most successful and widely adopted algorithms in robotics applications (Khan *et al.*, 2012). In fact, Sarsa and NQF can be considered to be modified versions of this scheme. The main difference between Sarsa and Q-learning is that the state-action value function is not necessarily updated with the policy that is being followed (Sutton and Barto, 1998):

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a' \in \mathcal{A}(s')} Q(s',a') - Q(s,a)\right]. \tag{4.13}$$

For this reason, Q-learning is known as an *off-policy* algorithm. Similarly to Sarsa, Q-learning is guaranteed to converge for discrete actions and states, a bounded reward variance, the use of a discount factor and a properly decaying learning rate (Jaakkola *et al.*, 1994).

Algorithm 3 shows the Q-learning scheme for exact, discrete states. Like for Sarsa, eligibility traces are not employed. Q-learning with linear function approximation for the state space is described in Section 4.5.5.

Algorithm 3: Q-learning: on-line, off-policy reinforcement learning algorithm with discrete, exact states, adapted from Sutton and Barto (1998).

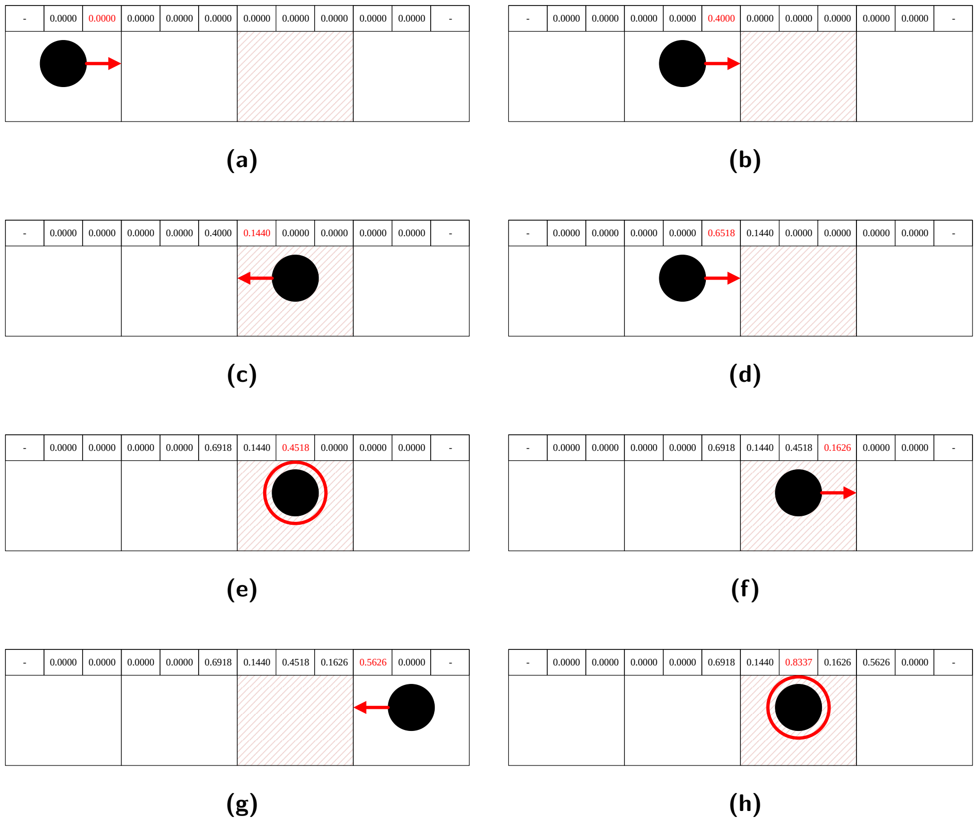
```
Input: $\mathcal{S}$, $\mathcal{A}$, $\alpha_0$, $N_\alpha$, $\gamma$, $\epsilon_0$, $N_\epsilon$
Output: $\pi$
initialize $Q(s, a)$ arbitrarily;
for each episode do
    initialize $\boldsymbol{N} \leftarrow \mathbf{0}$;
    initialize $s$;
    for each step in the episode do
        get $\epsilon$ with (4.10);
        choose $a$ given $s$ using an $\epsilon$-greedy policy with (4.9);
        take action $a$, observe $r$, $s'$;
        update $\boldsymbol{N}(s, a) \leftarrow \boldsymbol{N}(s, a) + 1$;
        get $\alpha$ with (4.12);
        update $Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a' \in \mathcal{A}(s)} Q(s', a') - Q(s, a)\right]$ ;
        update $s \leftarrow s'$ ;
    end
end
```

In the context of WEC control, Q-learning has been the first reinforcement learning algorithm that has been investigated. It has successfully been applied to the resistive and reactive control of a WEC, as can be seen in Chapter 6.

4.5.2.1 State-action value function example

Before moving on to the treatment of function approximation and other reinforcement learning algorithms, let us consider a simple example in order to fully understand the function of the reinforcement learning update of the state-action value function.

Let us consider a very simple, one-dimensional, grid-world navigation problem, where the robot is restricted to motions in only 4 cells arranged horizontally, as shown in Figure 4.3. At each step, the robot can decide to go left, right, or stay in the same cell, i.e. there are 3 actions in total. The leftmost and rightmost cells present only two actions to prevent the robot from exceeding the state space limits. The robot receives a reward of +1 whenever it moves to (or stays in) the third cell from the left. No ”living cost” penalty is awarded. The learning rate is set to 0.4 and the discount factor to 0.9.



(a) **(b)** **(c)** **(d)** **(e)** **(f)** **(g)** **(h)**

Figure 4.3: Position of the robot in the example one-dimensional grid world navigation problem at different steps. The red arrow indicates the selected action (or a ring if the robot is to stay in the same state). The dashed red square is associated with a reward of +1. The entries for the Q-table after the update in each step are displayed on the top of the cells.

From this description, it is clear that the problem presents 4 exact (or discrete) states and 3 actions (except at the end points). Hence, it is possible to represent the state-action value function as a matrix of size $(4, 3)$, where the entries $(1, 1)$ and $(4, 3)$ are void. Figure 4.3 shows how the Q-table is updated at the end of each step using (4.13) starting from $\boldsymbol{Q} = \mathbf{0}$ for an arbitrary selection of actions. The actions are represented by either the red arrows or the red ring (if the action is to stay in the same cell). The figure also displays the state-action values for each state-action pair at the top of each cell: the left, middle and right cells correspond to the actions move left, stay and move right, respectively. Each cell corresponds to a particular state. The updated state-action value at the end of the step is highlighted in red.

Now that the mechanism of temporal difference methods has been described with exact states and actions, the topic of the approximation of the state-action value is introduced. This technique can result in a significant computational saving for large state and action spaces.

4.5.3 Function approximation

So far, we have considered discrete, exact actions and states. Hence, the state-action value function could in fact be represented by a table (Sutton and Barto, 1998). Nevertheless, for control applications in engineering, continuous states and actions are more common (Busoniu *et al.*, 2010), e.g. the voltage or current of an electric motor. Function approximation can be used to describe Q as a function of the continuous actions and states. The approximate state-action value function is typically represented as $\hat{Q}$ and the notation is followed here as well. Most function approximation methods have been developed for discrete, exact actions. These schemes apply to bang-bang control types or whenever the control action corresponds to a step change in a particular variable at each time step (Busoniu *et al.*, 2010), e.g. a step change in torque of an electric motor. In fact, most systems nowadays present digital control strategies due to the low cost of micro-controllers, so that the assumption of discrete actions is not unrealistic. For this reason, only the state space is treated as continuous in this work, and we ignore more exotic schemes that fit the action space as well.

Not only does function approximation reduce the computational costs associated with the representation of large state and action spaces, but it also enables reinforcement learning algorithms to generalize for unseen states, with a possible decrease in convergence time (Geramifard *et al.*, 2013). The simplest approximation of a continuous state space is to use a number of discrete tiles (Sutton and Barto, 1998). The continuous space enclosed by the boundary of each tile corresponds to a particular discrete state. Such a procedure is also known as a tabular representation (Geramifard *et al.*, 2013), which presents a very low accuracy for most applications. Hence, a large number of tiles may be required to approximate the state space, which clearly results in a very large computational cost in the calculation and update of the state-action value function. Although it is possible to improve the computational efficiency of this particular strategy with either the use of hashing or the adoption of non-uniform tiles (Sutton and Barto, 1998), alternative methods may be more efficient. Here, we focus on linear features and neural networks for the function approximation of the state space.

4.5.3.1 Linear function approximation

The main advantages of linear function approximation (i.e. with linearly independent features) are its simple implementation and the ease of debugging and feature engineering (Lagoudakis and Parr, 2003). With this strategy (Geramifard *et al.*, 2013), the state-action value function is expressed in matrix notation as

$$Q(s,a) \approx \hat{Q}(s,a) = \boldsymbol{\phi}(s)^T \boldsymbol{\Theta}_{:,a}, \tag{4.14}$$

where $\boldsymbol{\Theta}$ is the weight matrix and $\boldsymbol{\phi}$ is the vector of arbitrary, linearly independent, usually non-linear basis functions, or features. The weight matrix presents a column for each discrete action (hence $|\mathcal{A}|$ columns in total), so that $\boldsymbol{\Theta}_{:,a}$ indicates the a^{th} column of $\boldsymbol{\Theta}$. $\boldsymbol{\Theta}$ and $\boldsymbol{\phi}$ have J rows. Usually, $J \ll |\mathcal{S}|$ so that the memory cost of a linear architecture is smaller than the exact representation, with $\mathcal{S}$ indicating the state space.

Although a wide range of basis functions are possible, in this work two types of feature are used: tabular and radial. As aforementioned, the tabular representation is the simplest and consists in assigning a separate weight for each state-action pair (Geramifard *et al.*, 2013). With this method, the j^{th} feature associated with the current state s_j is 1, while all other parameters zero:

$$\phi_j(s) = \begin{cases} 1 & \text{if } s = s_j \ , \quad (4.15) \\ 0 & \text{if } s \neq s_j \ . \quad (4.16) \end{cases}$$

For discrete states, this corresponds to the exact representation $Q(s,a)$, although its size is equal to the whole state-action space, i.e. $J = |\mathcal{S}|$. As a result, there is no gain in computational performance.

Gaussian radial basis functions (RBFs) (Moody and Darken, 1989) enable a continuous representation of the state space. These features have been widely applied to reinforcement learning problems (Busoniu *et al.*, 2010; Geramifard *et al.*, 2013). In RBFs, the feature activation decays continuously away from the state-action pair where the basis function is centred, s_j for the j^{th} RBF, spanning many discrete states (Geramifard *et al.*, 2013):

$$\phi_j(s) = \exp\left(-\frac{||s - s_j||^2}{2\mu_j}\right), \tag{4.17}$$

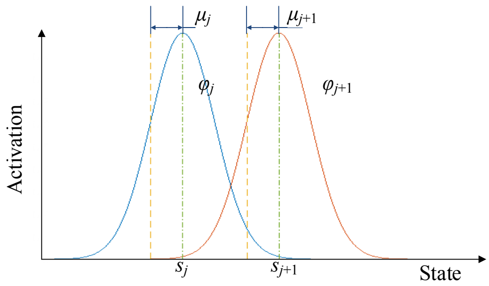


Figure 4.4: Activation function of radial basis functions in two dimensions as per (4.17).

where μ_j indicates the bandwidth of the j^{th} RBF. RBFs are shown graphically in two dimensions in Figure 4.4. A larger value of μ results in a flatter curve. Basis functions can present a different bandwidth for each direction, resulting in elliptical basis functions (Busoniu *et al.*, 2010). Since the output of RBFs decays to zero far away from its centre, the location of the centres has a strong influence on the accuracy and validity of the resulting representation (Geramifard *et al.*, 2013). For instance, Geramifard *et al.* (2013) have shown that a random distribution of RBFs can result in severe learning problems in the benchmark pole-cart problem. Different strategies to address this problem, including an adaptive location strategy, are discussed in Geramifard *et al.* (2013). Here, a uniform distribution of RBFs is employed. An additional bias term, which presents a value of 1 and a corresponding weight, is also included to provide the offset of the fit of the function approximation.

4.5.3.2 Neural function approximation

Recently, the application of non-linear basis functions has been proposed. Many strategies have been developed, including regression and machine learning techniques. An example is the use of regression trees in Ernst *et al.* (2005). Here, only neural networks have been considered based on NFQ described by Riedmiller (2005). Neural networks represent a powerful, non-linear tool that allows global approximation also for non-linear problems. Their main advantage is the capacity to generalize for unseen situations. However, they present also a major disadvantage: when updating the state-action value function, information from the current state-action pair may affect the prediction of other state-action pairs in an unforeseeable manner (Riedmiller, 2012), even overwriting previous information. As a result, feature engineering is much more complex than for linear basis functions. The method proposed by Riedmiller (2005) to overcome this issue is described in Section 4.5.6.

Here, a feed-forward multi-layer ANN with a single hidden layer with m neurons is considered, as shown in Figure 4.5. The output of the ANN can be obtained with forward

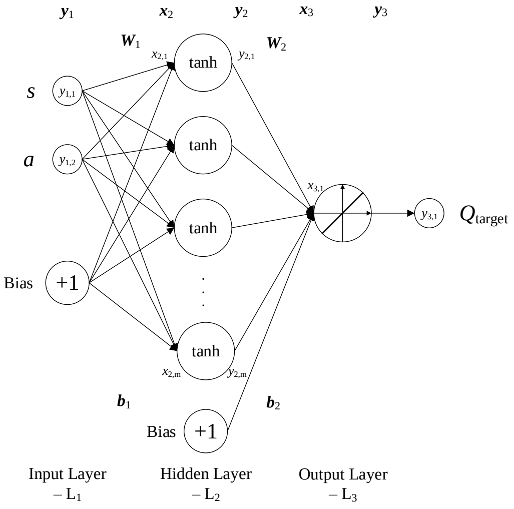


Figure 4.5: Schematic diagram of the feed-forward neural network used here for the function approximation of the state-action value (assuming one state and one action).

propagation, as described in Section A.2.3. Similarly, as explained in Section A.2.4, for a neural network learning occurs by tuning the weight matrices so that the network provides an accurate mapping between provided input and output data. To simplify the notation, here we denote the mapping provided by the neural between input $\boldsymbol{i}$ and output $\boldsymbol{o}$ as

$$\boldsymbol{o} = f(\boldsymbol{i}). \tag{4.18}$$

The neural network is then trained using input and output data with backward propagation as described in Section A.2.4. Although the Rprop algorithm (Riedmiller and Braun, 1993) is used in Riedmiller (2005), here we employ the efficient Levenberg-Marquardt scheme (Hagan and Menhaj, 1994) for training (Section A.2.4.2).

4.5.4 Sarsa with linear function approximation

The Sarsa algorithm is described for discrete states in Section 4.5.1. The more general form of the scheme with the inclusion of function approximation can be found in Algorithm 4. Only linear function approximation is considered here for simplicity. Even though different function approximation features are investigated, a tabular approach is used for the update of the exploration and learning rates. Hence, this means that the number of discrete states is $\mathcal{S}_{\mathrm{d}}$, each of which corresponds to a row of the $\boldsymbol{N}$ matrix. Thus, for each continuous state s, the corresponding closest discrete state s_{d} is found. This is specific to this particular application of reinforcement learning to the control of WECs, where only one episode is considered. For more standard applications, a greater number of episodes is used to teach the controller a particular task, e.g. acrobatic manoeuvres for helicopters (Abbeel, 2008). Therefore, it is usually preferred to update the exploration and learning rates based on the number of episodes (Busoniu *et al.*, 2010; Geramifard *et al.*, 2013).

Algorithm 4: Sarsa: on-line, on-policy reinforcement learning algorithm with linear function approximation, adapted from Sutton and Barto (1998).

```
Input: $\mathcal{A}$, $\alpha_0$, $N_\alpha$, $\gamma$, $\epsilon_0$, $N_\epsilon$, $\mathcal{S}_\mathrm{d}$, $\boldsymbol{\phi}$
Output: $\pi$
initialize $\boldsymbol{\Theta}$ arbitrarily;
for each episode do
    initialize $\boldsymbol{N} \leftarrow \mathbf{0}$;
    initialize $s$;
    for all $a \in \mathcal{A}(s)$ do
        get $\hat{Q}(s,a) = \boldsymbol{\phi}(s)^T \boldsymbol{\Theta}_{:,a}$ with (4.14);
    end
    get $\epsilon$ with (4.10);
    choose $a$ given $s$ using an $\epsilon$-greedy policy with (4.9);
    for each step in the episode do
        get the discrete state $s_\mathrm{d}$ from $s$;
        update $\boldsymbol{N}(s_\mathrm{d}, a) \leftarrow \boldsymbol{N}(s_\mathrm{d}, a) + 1$;
        get $\epsilon$ with (4.10);
        get $\alpha$ with (4.12);
        take action $a$, observe $r$, $s'$;
        initialize $\delta = r - \hat{Q}(s,a)$;
        for all $a \in \mathcal{A}(s')$ do
            get $\hat{Q}(s',a) = \boldsymbol{\phi}(s')^T \boldsymbol{\Theta}_{:,a}$ with (4.14);
        end
        choose $a'$ given $s'$ using an $\epsilon$-greedy policy (4.9);
        update $\delta \leftarrow \delta + \gamma \hat{Q}(s',a')$;
        update $\boldsymbol{\Theta}_{s,a} \leftarrow \boldsymbol{\Theta}_{s,a} + \alpha\delta$ ;
        update $s \leftarrow s'$ ;
        update $a \leftarrow a'$;
    end
end
```

4.5.5 Q-learning with linear function approximation

Q-learning for discrete states and actions is described in Section 4.5.2. Algorithm 5 presents the more general version for linear function approximation. In this case, as for Sarsa, discrete states are used to update the learning and exploration rates.

Algorithm 5: Q-learning: on-line, off-policy reinforcement learning algorithm with linear function approximation, adapted from Sutton and Barto (1998).

```
Input: $\mathcal{A}$, $\alpha_0$, $N_\alpha$, $\gamma$, $\epsilon_0$, $N_\epsilon$, $\mathcal{S}_\mathrm{d}$, $\boldsymbol{\phi}$
Output: $\pi$
initialize $\boldsymbol{\Theta}$ arbitrarily;
for each episode do
    initialize $\boldsymbol{N} \leftarrow \mathbf{0}$;
    initialize $s$;
    for each step in the episode do
        for all $a \in \mathcal{A}(s)$ do
            get $\hat{Q}(s,a) = \boldsymbol{\phi}(s)^T \boldsymbol{\Theta}_{:,a}$ with (4.14);
        end
        get $\epsilon$ with (4.10);
        choose $a$ given $s$ using an $\epsilon$-greedy policy with (4.9);
        get discrete state $s_\mathrm{d}$ from $s$;
        update $\boldsymbol{N}(s_\mathrm{d}, a) \leftarrow \boldsymbol{N}(s_\mathrm{d}, a) + 1$;
        get $\alpha$ with (4.12);
        take action $a$, observe $r$, $s'$;
        initialize $\delta = r - \hat{Q}(s,a)$;
        for all $a \in \mathcal{A}(s')$ do
            get $\hat{Q}(s',a) = \boldsymbol{\phi}(s')^T \boldsymbol{\Theta}_{:,a}$ with (4.9);
        end
        update $\delta \leftarrow \delta + \gamma \max_{a' \in \mathcal{A}(s)} \hat{Q}(s',a')$;
        update $\boldsymbol{\Theta}_{s,a} \leftarrow \boldsymbol{\Theta}_{s,a} + \alpha\delta$ ;
        update $s \leftarrow s'$ ;
    end
end
```

4.5.6 Neural fitted Q-iteration

NFQ was originally proposed by Riedmiller (2005). The scheme was developed to combine Q-learning with neural function approximation so as to exploit the advantages of both approaches: a strong, efficient learning algorithm and a tool that is able to fit non-linear functions accurately. According to the Q-learning state-action value function update in (4.13), the target of the neural network (i.e. the output) can be set to

$$\hat{Q}_{\text{target}}(s,a) = r + \gamma \max_{a' \in \mathcal{A}(s')} \hat{Q}(s',a'). \tag{4.19}$$

As a result, the output squared error of the neural network is given by

$$e = \left(\hat{Q}(s,a) - \hat{Q}_{\text{target}}(s,a)\right)^2, \tag{4.20}$$

which can be then backward propagated as described in Chapter **??**. In (4.19), $\hat{Q}$ is denoted as approximate, as it is computed by the ANN.

However, in order to solve the problem caused by the training of neural networks on-line, a batch-mode algorithm was developed instead. This means that the state-action value function is updated off-line at regular intervals using samples of the form (s,a,r,s') that are stored at each step from the interactions with the environment. In particular, the whole set of transition experiences (i.e. all past samples) and the Levenberg-Marquardt algorithm are used for the training of the neural network weights. This procedure is repeated for a specified number of epochs, $k_{\max}$. This approach has been found to work well, and is more computationally efficient than setting a limit for the error (Riedmiller, 2012). In a practical application, a finite number of samples can be stored due to memory requirements and the computational cost associated with training the neural network. It will be important to ensure a broad range of samples is maintained in order to produce a network of sufficient quality.

Neural fitted Q-iteration is summarized in Algorithm 6. The notation $\boldsymbol{S}_{:,j}$ indicates the j^{th} column vector of the list of samples $\boldsymbol{S}$. The condition for the update of the weights of the neural network usually corresponds to the collection of a specified number of samples.

Neural fitted Q-iteration has not been applied to the control of WECs, but rather to the control of tidal turbines.

Algorithm 6: Neural fitted Q-iteration: off-line, off-policy reinforcement learning algorithm with neural function approximation, adapted from Riedmiller (2005).

```
Input: $\mathcal{A}$, $N_\alpha$, $\gamma$, $\epsilon_0$, $N_\epsilon$, $\mathcal{S}_\mathrm{d}$
Output: $\pi$
initialize $\boldsymbol{W}$ (neural network weights) randomly;
initialize $\boldsymbol{S} \leftarrow []$ (empty samples list);
get number of actions $|\mathcal{A}|$;
for each episode do
    for each step in the episode do
        observe current continuous state $s$;
        get corresponding discrete state $s_\mathrm{d}$;
        update exploration rate $\epsilon$ with (4.10);
        for each action $a$ in $\mathcal{A}$ do
            get $\hat{Q}(s,a) \approx f(s,a)$ with (4.18);
        end
        select an action $a$ with an $\epsilon$-greedy policy with (4.9);
        apply action $a$ and observe reward $r$ and new state $s'$;
        store $\boldsymbol{S} \leftarrow [\boldsymbol{S}; (s,a,r,s')]$;
        if condition met for off-line neural network update then
            get the number of samples $n_\mathrm{s}$;
            initialize $\hat{\boldsymbol{Q}}_\mathrm{target} = 0$ with size $(n_\mathrm{s}, |\mathcal{A}|)$;
            for $k = 1 : k_\mathrm{max}$ do
                for each sample $i$ in $\boldsymbol{S}$ do
                    for each action $a$ in $\mathcal{A}$ do
                        get $\hat{Q}(s',a') \approx f\left(\boldsymbol{S}(i,4), a'\right)$ with (4.18);
                    end
                    get $\hat{\boldsymbol{Q}}_\mathrm{target}(i) = \boldsymbol{S}(i,3) + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s',a')$ with (4.19);
                end
                train neural network as in Hagan and Menhaj (1994) with input $\boldsymbol{S}(:,1)$
                  and $\boldsymbol{S}(:,2)$ and output $\hat{\boldsymbol{Q}}_\mathrm{target}$;
            end
        end
    end
end
```

4.5.7 Least-squares policy iteration

Least-squares policy iteration is an off-line, on-policy reinforcement learning scheme developed by Lagoudakis and Parr (2003). Its inspiration is taken from *policy iteration*, a technique to find the optimal policy for any Markov decision process (Howard, 1960). In a reinforcement learning framework, the underlying Markov decision process is not available. Thus, the algorithm must rely on information coming from direct interactions with the environment. At each time step, observations are stored as *samples* of the form (s, a, r, s') (Lagoudakis and Parr, 2003). As before, s indicates the current state, a the action taken by the controller, r is the observed reward and s' indicates the new state the agent lands into.

Least-squares policy iteration is an approximate policy-iteration algorithm that learns decision policies from the stored samples (Lagoudakis and Parr, 2003). The state-action value, or Q-value, is approximated with linear features and calculated on demand. This means that for any desired state s, the Q-value is computed for all actions using the stored parameters of the approximation. The greedy action is then selected. *Least-squares temporal difference Q* (LSTDQ) (Lagoudakis and Parr, 2003) is used to update the policy using the stored samples. In the following subsections, these steps will be described in detail.

4.5.7.1 Policy iteration

As opposed to the other temporal difference methods described above, in policy iteration the policy is represented with a separate memory structure independent of the value function (Sutton and Barto, 1998). The optimal policy is found by alternating iteratively between a stage of *policy evaluation*, where the Q-value is estimated for the current policy from the linear system of the Bellman equations, and a stage of *policy improvement*, where the policy is updated to the policy that optimizes the action-value (Lagoudakis and Parr, 2003). In this particular application, the policy is improved by using the ϵ-greedy policy previously described. The two steps of the process are repeated until there is no longer a change in the policy, which has fully converged to the optimal policy. Policy evaluation is also referred to as the *critic* and policy improvement as the *actor* (Lagoudakis and Parr, 2003). Therefore, policy iteration strategies are known as actor-critic architectures (Sutton and Barto, 1998).

Policy iteration is guaranteed to converge to the optimal policy only for a tabular representation of the state-action value function, exact solution of the Bellman equations and a tabular representation of the policy (Lagoudakis and Parr, 2003). The associated computational cost becomes excessive for large state and action spaces, so that approximation methods are usually employed. In particular, two approximations are typically made (Lagoudakis and Parr, 2003):

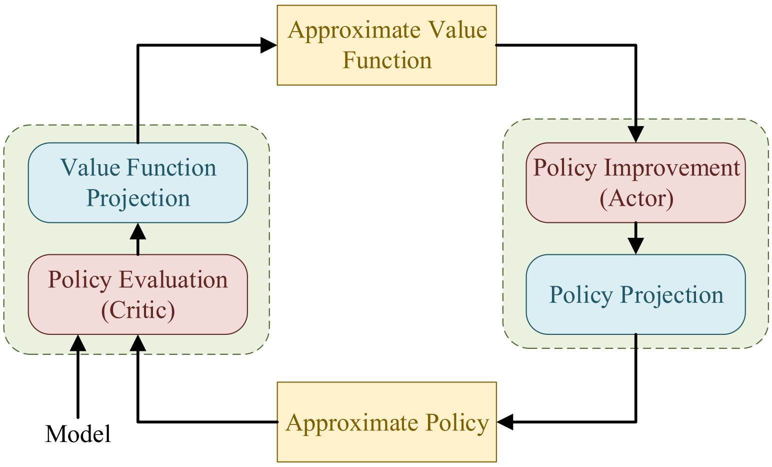

Approximate Value Function
Value Function Projection
Policy Evaluation (Critic)
Policy Improvement (Actor)
Policy Projection
Model
Approximate Policy

Figure 4.6: Diagram of approximate policy iteration, adapted from Lagoudakis and Parr (2003).

- The exact representation of the value function $Q^\pi(s, a)$ (for policy π, state s and action a) is replaced by the parametric function $\hat{Q}^\pi(s, a, w)$, where w indicates the adjustable weights of the approximator.
- The exact policy $\pi(s)$ is replaced by the approximate representation $\hat{\pi}(s, \theta)$, where θ corresponds to the adjustable parameters of the representation.

As only the parameters need to be stored, the memory requirements are much smaller for *approximate policy iteration* than for exact policy iteration. The policy evaluation and the action-value estimation (or projection) are merged into a single phase, while policy improvement and the policy projection are blended into the second phase. Approximate policy iteration is shown graphically in Figure 4.6.

The soundness of the approximation strategy of policy iteration is supported by the findings by Bertsekas and Tsitsiklis (1996) and Munos (2003). These studies proved that approximate policy iteration converges to the optimal policy if the errors in the approximation of the policy and Q-value are bounded and decrease to zero.

4.5.7.2 Least-squares fixed-point approximation

The state value function can be approximated through the Bellman residual minimizing approach (Schweitzer and Seidmann, 1985), based on the Bellman equations in (4.4). Although the Bellman residual minimizing technique is more stable and predictable, a different method, known as least-squares fixed-point, is more suitable for the approximation of the state-action value function in the context of learning (Munos, 2003). In particular, only a generative model of the Markov decision process can produce the ”doubled” samples necessary for the former approach (Lagoudakis and Parr, 2003).

Furthermore, the policies obtained from least-squares fixed-point approximation have been found to be superior in practice (Lagoudakis and Parr, 2003). For these reasons, only this method is presented here.

The least-squares fixed-point scheme has been developed by Bradtke *et al.* (1996) for learning the state-action value function from samples. This method is based on the observation that the state-action value function Q^{π} of policy π is a fixed point of the Bellman operator T_{π} (Lagoudakis and Parr, 2003):

$$T_{\pi} Q^{\pi} = Q^{\pi}, \tag{4.21}$$

which is identical to (4.4). Therefore, a good approximation for the value function would be a fixed point under the Bellman operator, which must lie in the space spanned by the basis functions (Lagoudakis and Parr, 2003). Although Q^{π} lies there by definition, this may not be true for $T_{\pi} Q^{\pi}$, which must be projected. Lagoudakis and Parr (2003) consider an orthogonal projection that minimizes the L_2 norm: $\left(\mathbf{\Phi}^T (\mathbf{\Phi}\mathbf{\Phi})^{-1} \mathbf{\Phi}^T\right)$. The problem reduces to the search for the approximate value function $\hat{Q}^{\pi}$ that is invariant under one application of the Bellman operator T_{π} followed by the orthogonal projection (Lagoudakis and Parr, 2003):

$$\hat{Q}^{\pi} = \left(\mathbf{\Phi}^T (\mathbf{\Phi}\mathbf{\Phi})^{-1} \mathbf{\Phi}^T\right) \left(T_{\pi} \hat{Q}^{\pi}\right), \tag{4.22}$$

$$\hat{Q}^{\pi} = \left(\mathbf{\Phi}^T (\mathbf{\Phi}\mathbf{\Phi})^{-1} \mathbf{\Phi}^T\right) \left(\mathcal{R} + \gamma \boldsymbol{P} \mathbf{\Pi}_{\pi} \hat{Q}^{\pi}\right). \tag{4.23}$$

Since linearly independent features are used for the function approximation, the column space of $\mathbf{\Phi}$ is well defined. Rearranging (4.23), it is possible to express the problem as the solution of a $(K \times K)$ linear system, with K being the total number of basis functions (Lagoudakis and Parr, 2003):

$$\mathbf{\Phi}^T (\mathbf{\Phi} - \gamma \boldsymbol{P} \mathbf{\Pi}_{\pi} \mathbf{\Phi}) w^{\pi} = \mathbf{\Phi}^T \mathcal{R}. \tag{4.24}$$

For a finite number of any γ values and for any $\mathbf{\Pi}_{\pi}$, the solution to the system, which is known as least-squares fixed-point approximation, is guaranteed to exist (Koller and Parr, 2000):

$$w^{\pi} = \left(\mathbf{\Phi}^T (\mathbf{\Phi} - \gamma \boldsymbol{P} \mathbf{\Pi}_{\pi} \mathbf{\Phi})\right)^{-1} \mathbf{\Phi}^T \mathcal{R}. \tag{4.25}$$

In order to control the distribution of the approximation error, a weighted projection is typically employed. Defining μ as the probability distribution over (s, a) and $\mathbf{\Delta}_{\mu}$ as the diagonal matrix with the projection weights $\mu(s, a)$, the weighted least-squares fixed-point approximation is (Lagoudakis and Parr, 2003)

$$w^{\pi} = \left(\mathbf{\Phi}^T \mathbf{\Delta}_{\mu} (\mathbf{\Phi} - \gamma \boldsymbol{P} \mathbf{\Pi}_{\pi} \mathbf{\Phi})\right)^{-1} \mathbf{\Phi}^T \mathbf{\Delta}_{\mu} \mathcal{R}. \tag{4.26}$$

4.5.7.3 Least-squares temporal difference learning for the state-action value function

With temporal difference methods, the model of the environment is not available. Hence, the weighted least-squares fixed-point approximation must be learned from samples. Taking into account that there are K linearly independent basis functions, this problem reduces to learning the parameters w^{π} of $\hat{Q}^{\pi} = \mathbf{\Phi} w^{\pi}$ (Lagoudakis and Parr, 2003). From (4.26), the exact values for w^{π} are computed by solving the following linear system of equations (Lagoudakis and Parr, 2003):

$$\boldsymbol{A} w^{\pi} = b, \tag{4.27}$$

where

$$\boldsymbol{A} = \left(\mathbf{\Phi}^T \mathbf{\Delta}_{\mu} (\mathbf{\Phi} - \gamma \boldsymbol{P} \mathbf{\Pi}_{\pi} \mathbf{\Phi})\right), \tag{4.28}$$

$$b = \mathbf{\Phi}^T \mathbf{\Delta}_{\mu} \mathcal{R}, \tag{4.29}$$

and μ is a probability distribution over $(\mathcal{S} \times \mathcal{A})$ that describes the weight of the projection.

In temporal difference methods, $\boldsymbol{A}$ and b are not known a-priori, but rather must be learned from samples. Denoting learned values with $\tilde{}$, considering samples of the form (s, a, r, s') and assuming that the distribution of the samples matches μ, Lagoudakis and Parr (2003) have shown that for L samples the learned values of $\boldsymbol{A}$ and b can be expressed as:

$$\tilde{\boldsymbol{A}} = \frac{1}{L} \left(\tilde{\mathbf{\Phi}}^T (\tilde{\mathbf{\Phi}} - \gamma \widetilde{\boldsymbol{P} \mathbf{\Pi}_{\pi} \mathbf{\Phi}})\right), \tag{4.30}$$

$$\tilde{b} = \tilde{\mathbf{\Phi}}^T \tilde{\mathcal{R}}, \tag{4.31}$$

where

$$\tilde{\mathbf{\Phi}} = \begin{bmatrix} \boldsymbol{\phi}(s_1, a_1)^T & \ldots & \boldsymbol{\phi}(s_l, a_l)^T & \ldots & \boldsymbol{\phi}(s_L, a_L)^T \end{bmatrix}^T, \tag{4.32}$$

$$\widetilde{\boldsymbol{P} \mathbf{\Pi}_{\pi} \mathbf{\Phi}} = \begin{bmatrix} \boldsymbol{\phi}({s'}_1, \pi({s'}_1))^T & \ldots & \boldsymbol{\phi}({s'}_i, \pi({s'}_l)^T & \ldots & \boldsymbol{\phi}({s'}_L, \pi({s'}_L)^T \end{bmatrix}^T, \tag{4.33}$$

$$\tilde{\mathcal{R}} = \begin{bmatrix} r_1 & \ldots & r_l & \ldots & r_L \end{bmatrix}^T. \tag{4.34}$$

In the limit of an infinite number of samples, $\tilde{\boldsymbol{A}}$ and $\tilde{b}$ converge to $\boldsymbol{A}$ and b. In this particular case, an ϵ-greedy strategy is employed. Hence, $\pi(s)$ is given by the evaluation of the state-value function for all available actions for the current state and the selection of the one that results in the maximum value, unless a random action is chosen.

A problem with (4.27) is that $\tilde{\boldsymbol{A}}$ needs to be inverted. However, the matrix will not be full rank until a sufficient number of samples has been collected. Lagoudakis and Parr (2003) have proposed the adoption of recursive least-squares techniques to compute the

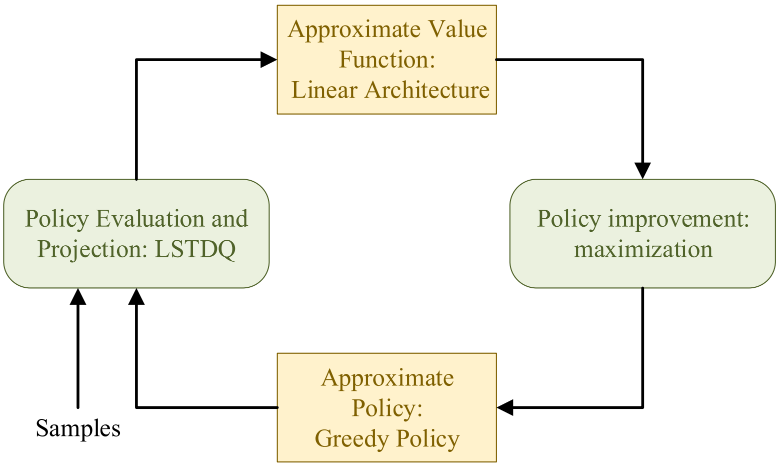

Approximate Value Function: Linear Architecture
Policy Evaluation and Projection: LSTDQ
Policy improvement: maximization
Approximate Policy: Greedy Policy
Samples

Figure 4.7: Diagram of the LSPI algorithm, adapted from Lagoudakis and Parr (2003).

inverse of $\tilde{\boldsymbol{A}}$ recursively and more information can be found in their work.

4.5.7.4 Least-squares policy iteration algorithm

At this point, using the knowledge of linear function approximation, approximate policy iteration, least-squares fixed-point approximation and least-squares temporal difference learning for the state-action value function, it is possible to express LSPI as in Algorithm 7. In addition, Figure 4.7 summarizes LSPI graphically.

In a WEC control context, least-squares policy iteration has been applied to the resistive and reactive control of a point absorber, as can be seen in Chapter 6. In particular, the performance of LSPI has been assessed against Q-learning and Sarsa.

4.6 Benchmark problems

The performance of the Sarsa, Q-learning and LSPI algorithms has been assessed using two established benchmark cases: a grid-world navigation problem and the control of an inverted pendulum on a cart. Both discrete states (i.e. tabular features) and RBFs are used to approximate the state space for all algorithms. Monte-Carlo methods and neural fitted Q-iteration have been neglected from these studies, since they have been analysed only peripherally in this project. Nevertheless, it is possible to find a successful implementation of neural fitted Q-iteration for the inverted pendulum problem in Riedmiller (2005).

Algorithm 7: Least-squares policy iteration: off-line, on-policy reinforcement learning algorithm with linear function approximation, adapted from Lagoudakis and Parr (2003).

```
Input: $\mathcal{A}$, $\alpha_0$, $N_\alpha$, $\gamma$, $\epsilon_0$, $N_\epsilon$, $\mathcal{S}_\mathrm{d}$, $\boldsymbol{\phi}$, $\delta$
Output: $\pi$, $\boldsymbol{\Theta}$
initialize $\boldsymbol{\Theta}$ arbitrarily;
initialize $\boldsymbol{\Theta}_0 = \mathbf{0}$;
initialize $\boldsymbol{S} \leftarrow []$ (empty samples list);
get number of actions $|\mathcal{A}|$;
for each episode do
    for each step in the episode do
        observe current continuous state $s$;
        get corresponding discrete state $s_\mathrm{d}$;
        update exploration rate $\epsilon$ with (4.10);
        for each action $a$ in $\mathcal{A}$ do
            $\hat{Q}(s,a) = \boldsymbol{\phi}(s)^T \boldsymbol{\Theta}_{:,a}$;
        end
        select an $\epsilon$-greedy action $a$ with (4.9);
        apply action $a$;
        observe reward $r$;
        observe new continuous state $s$;
        store $\boldsymbol{S} \leftarrow [\boldsymbol{S}; (s,a,r,s')]$;
        if condition met for off-line weight matrix update then
            $\boldsymbol{\Theta}_1 \leftarrow \boldsymbol{\Theta}_0$;
            while $||\boldsymbol{\Theta} - \boldsymbol{\Theta}_1|| \geq \delta$ do
                $\boldsymbol{\Theta}_1 \leftarrow \boldsymbol{\Theta}$;
                $\tilde{\boldsymbol{A}} \leftarrow \mathbf{0}$;
                $\boldsymbol{b} \leftarrow \mathbf{0}$;
                for each sample $(s,a,r,s') \in \boldsymbol{S}$ do
                    $\tilde{\boldsymbol{A}} \leftarrow \tilde{\boldsymbol{A}} + \boldsymbol{\phi}(s)\left(\boldsymbol{\phi}(s) - \gamma\boldsymbol{\phi}\left(s', \pi(s')\right)\right)^T$;
                    $\tilde{\boldsymbol{b}} \leftarrow \tilde{\boldsymbol{b}} + \boldsymbol{\phi}(s) r$;
                    for each action $a \in \mathcal{A}$ do
                        $\boldsymbol{\Theta}_{:,a} \leftarrow \tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{b}}$;
                    end
                end
            end
        end
    end
end
```

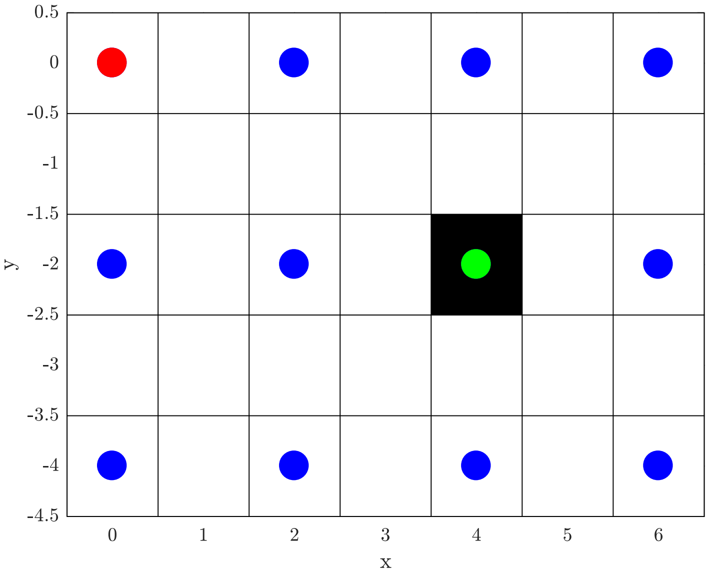


Figure 4.8: Representation of the analysed grid world. The starting position of the agent is represented by the red dot, while the black cell and green dot indicate the desired final position.

4.6.1 Grid-world navigation

Grid-world problems are some of the simplest established test cases for reinforcement learning algorithms and examples can be found in Sutton and Barto (1998), Busoniu *et al.* (2010) and Geramifard *et al.* (2013). In these situations, the reinforcement learning agent is placed in a two-dimensional, maze world with discrete states. At each step, it can take an action that results in a change of state and a movement on the grid from one cell to a neighbouring one. Some small uncertainty may be applied to the action implementation by adding a random noise element to the action selection process. The aim of the task consists in having the agent learn how to navigate from the starting position to a fixed cell in the grid, which is associated with a positive reward. Usually, a number of obstacles can be found in the form of walls or, worse, pits, which result in a penalty. An additional smaller penalty may be applied at each step due to ”living” costs, whose function is to speed up the selection of the shortest path. Despite their simplicity, grid-world are central to the discrete version of WEC control based on reinforcement learning treated in Chapter 6.

Here, a very basic grid-world problem is considered, which is shown in Figure 4.8. The blue dots indicate the position of the centres of the RBFs, with $\mu = 2$. The green and red dots also correspond to two additional centres. Furthermore, the red dot indicates the cell from which the agent starts in each episode. The green dot indicates the desired ultimate position to be reached, with the black square indicating that the episode is completed once the agent reaches this point and decides to stay in the same spot in the next step. Before completion, a reward of +1 is returned to the agent for reaching the desired goal. No living penalty cost is applied. Similarly, no obstacles are included, although the agent cannot go beyond the grid-world boundaries. Hence, the action selection is constrained on the limit of the state space. Otherwise, 5 actions are available to the agent: either take one step up, right, down or left, or stay in the same spot in the next step as well. No random noise is applied to the action selection for simplicity, which results in a deterministic policy. The exploration and learning rates present a simpler decay formulation than (4.10) and (4.12), respectively. The learning and exploration rates are initialized as $\alpha \leftarrow 0.4$ and $\epsilon \leftarrow 0.6$, respectively. At the end of each episode, both rates are updated as

$$\alpha \leftarrow 0.99\alpha \text{ and } \epsilon \leftarrow 0.99\epsilon, \text{ respectively.} \tag{4.35}$$

The discount factor is set to $\gamma = 0.95$.

For all algorithms, 1000 episodes are run for 100 different seed values to the random number generator. For LSPI, the weights of the function approximation are updated only once every 50 episodes. Note that this is different from the application to the control of WECs, where the update is performed after a specified number of steps rather than episodes. As shown in Figure 4.9, all algorithms converge towards the optimal policy in less than 1000 episodes, i.e. such that the agent completes the episode in 8 steps (including lingering in the final position). The figure displays the curve of the mean number of steps with episode number over all 100 repeats with 95% confidence limits as well as the best and worst curves over all cases (i.e. the lowest and highest number of steps, respectively). It is very important to notice the different scales of the x-axis for the different strategies. From Figure 4.9, it is clear that the on-line Sarsa and Q-learning schemes with discrete states perform best for the grid-world problem, due to the small number of *discrete* states. Furthermore, the associated computational cost (not displayed) is much lower. LSPI with tabular features takes longer to converge, since the weights are updated only once every 50 episodes. Despite the discrete nature of the states, even the schemes with function approximation are able to learn the optimal policy. However, this may be due to the location of the centres of the RBFs, with one lying on the final state.

In this simple problem, it is possible to visualize the optimal policy in Figure 4.10, which shows the policy to which Sarsa with discrete states converges. This is achieved by displaying the direction associated with the action that results in the maximum state-action value for each discrete state with arrows. In fact, in this case the algorithms converge towards different optimal policies for the various seeds to the random number generator due to the exploration process. However, all policies are optimal in that they

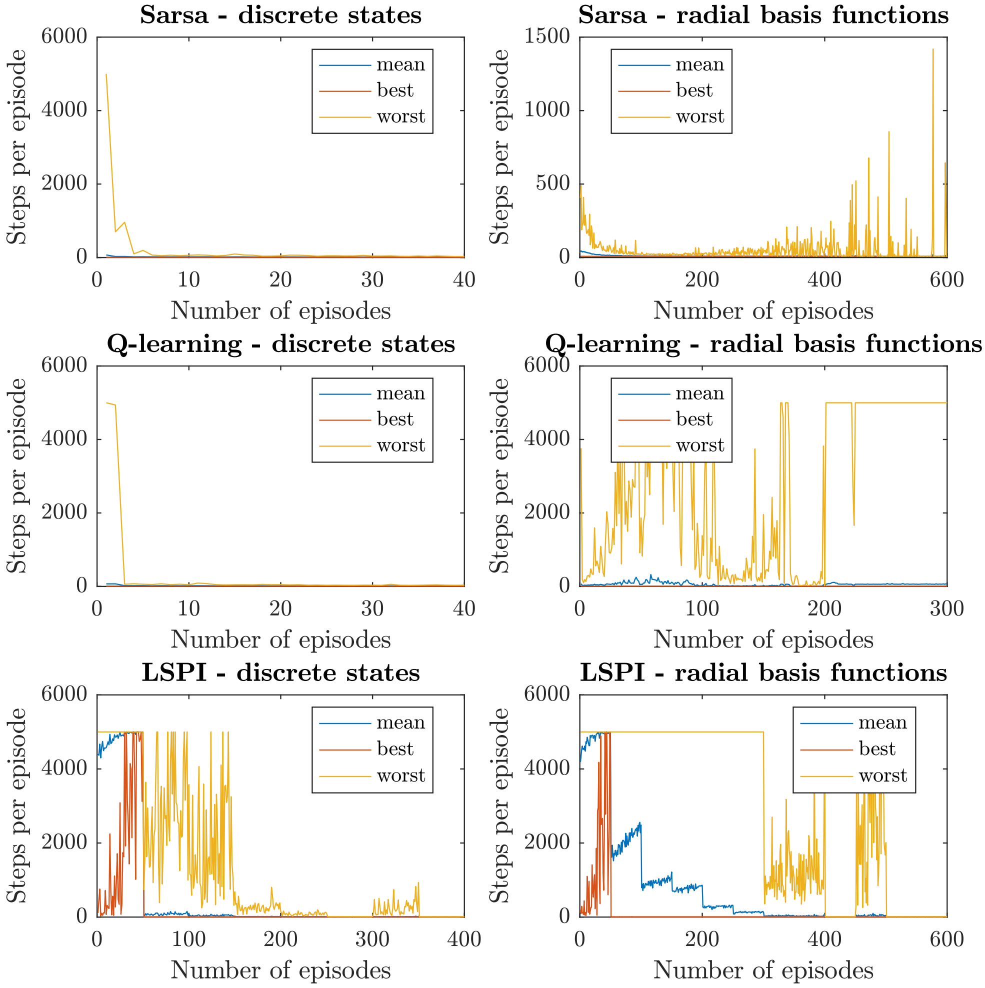


Figure 4.9: Plots of the number of steps required to reach the optimal point versus number of episodes for the three considered reinforcement learning algorithms with and without function approximation for the navigation in the grid world.

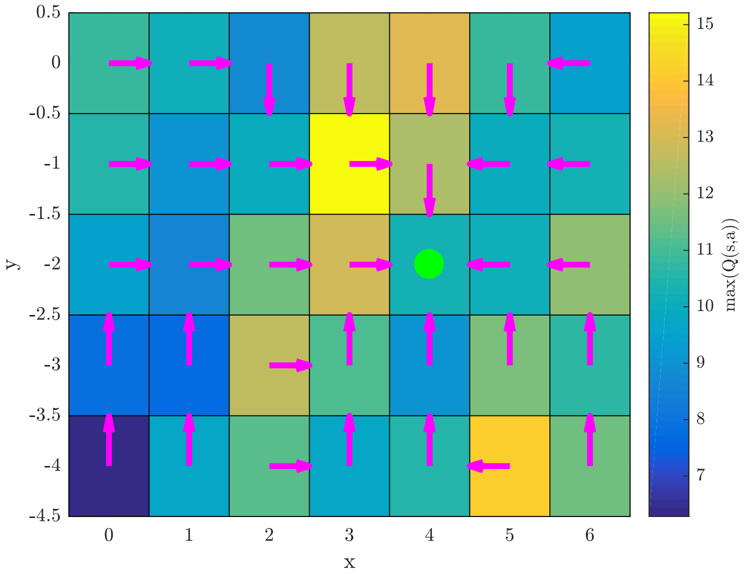


Figure 4.10: Policy selected by the Sarsa scheme with discrete states for the first seed value.

present 8 steps before completing the episode.

4.6.2 Control of an inverted pendulum on a cart

The second reinforcement learning task consists in the control of an inverted pendulum on a cart, which can be seen in Figure 4.11. No model is provided to the reinforcement learning algorithm and the agent needs to learn how to balance the pole by applying a fixed force onto the cart. This is also a renown benchmark problem for the assessment of reinforcement learning schemes (Busoniu *et al.*, 2010). Here, we consider the specific variant treated in Lagoudakis and Parr (2003), Riedmiller (2005) and Geramifard *et al.* (2013). This case study is considered to show the importance of function approximation in standard control tasks.

The motion of the mass on the top end of the pole can be expressed by the following equation of motion (Lagoudakis and Parr, 2003):

$$\ddot{\theta}(t) = \frac{g \sin(\theta(t)) - \alpha m l \left(\dot{\theta}(t)\right)^2 \sin(2\theta(t))/2 - \alpha \cos(\theta(t))\, u}{4l/3 - \alpha m l \cos^2(\theta(t))}, \tag{4.36}$$

where $g = 9.81$ m/s^2 is the gravitational acceleration, $m = 2$ kg the mass at the end of the pole, which is assumed to be massless and with length $l = 0.5$ m, and $\alpha = 1/(m + M)$, with $M = 8$ kg being the mass of the cart. In (4.36), θ is the instantaneous angle between the pole and the vertical line, as shown in Figure 4.11,

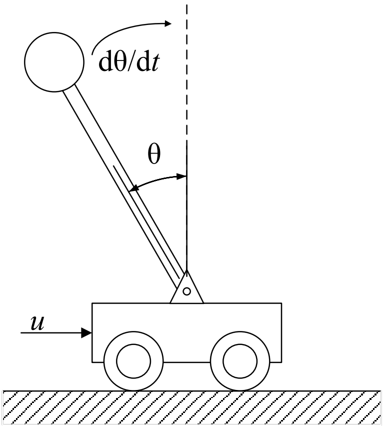


Figure 4.11: Control of an inverted pendulum on a cart.

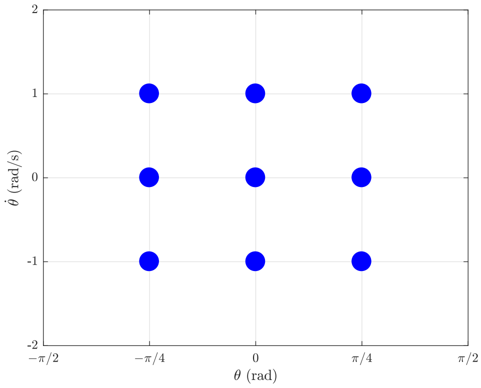


Figure 4.12: Position of the centres of the RBFs used in the reinforcement learning formulation of the control of the inverted pendulum. The position of the centres has been selected to match that proposed by Geramifard *et al.* (2013).

while u is the control force applied to the cart. In order to solve (4.36) numerically, a second state variable is introduced: $\Theta(t) = \dot{\theta}(t)$. Hence, (4.36) is rearranged into the following system of equations:

$$\dot{\theta}(t) = \Theta(t), \tag{4.37a}$$

$$\dot{\Theta}(t) = \frac{g\sin\left(\theta(t)\right) - \alpha ml\left(\Theta(t)\right)^2 \sin\left(2\theta(t)\right)/2 - \alpha\cos\left(\theta(t)\right)u}{4l/3 - \alpha ml\cos^2\left(\theta(t)\right)}. \tag{4.37b}$$

Equations (4.37a) and (4.37b) have been solved using a fourth-order Runge-Kutta scheme (Süli and Mayers, 2003) and a time step of 0.1 s.

For this reinforcement learning problem, each episode finishes when the pole falls, i.e. $|\theta| \geq \pi/2$ rad. When this happens, a penalty of -1 is returned; otherwise a return of 0 is observed. Hence, the agent should learn to balance the pole for as long as possible. The action space consists of three actions: apply a control force of −50 N, 0 N, or +50 N on the cart. In this case, the actions present no constraints; however, a uniform, random noise in $[-10, 10]$ N is added to the selected action. The state-space is represented by the two state variables, namely the angular displacement and velocity of the pole (θ and Θ, respectively). Here, only 9 features are used for each action for both tabular and radial basis functions in addition to a bias feature for each action, for a total of 30 basis functions for each algorithm. A regular distribution of the features is employed, as shown in Figure 4.12. The dots in Figure 4.12 correspond to the centres of either the tiles or the RBFs for tabular and radial features, respectively. A value of $\mu = 1$ is employed. The same exploration rate, learning rate and discount factor formulations as for the grid-world problem in Section 4.6.1 are adopted. Each episode is run for a maximum of 3000 steps, corresponding to 5 minutes. The control is deemed to be successful if the pole does not fall for the whole episode duration. For all algorithms, 1000 episodes have been investigated, which have been repeated 100 times with a different value of the seed to the random number generator. As for the grid world problem, the weights of the function approximation are updated only once every 50 episodes for the LSPI algorithm. The learning behaviour of the analysed reinforcement learning schemes can be seen in Figure 4.13. The figure displays the curve of the mean number of steps with episode number over all 100 repeats with 95% confidence limits as well as the best and worst curves over all cases (i.e. the highest and lowest number of steps, respectively).

Note that on purpose the same number of features is used for the tabular and radial basis functions, no experience replay is applied to the Q-learning and Sarsa algorithms and no pre-training data is provided to LSPI as opposed to the work presented by Lagoudakis and Parr (2003) and Geramifard *et al.* (2013). From the analysis of Figure 4.13, it is possible to make the following observations:

- the decay of the exploration and learning rates is too abrupt: the exploitative action is always selected for a number of episodes greater than 400.
- the number of tabular features is too small. Geramifard *et al.* (2013) have shown that 400 discrete states are required to ensure the pole is balanced for at least 5 minutes.
- LSPI with discrete sates performs worse than Q-learning and Sarsa, as observed in Section 4.6.1 for the grid world navigation problem.
- Experience replay is fundamental in ensuring convergence of Q-learning and Sarsa with function approximation, as shown by Lagoudakis and Parr (2003) for Q-learning.
- LSPI with RBFs is the only algorithm that manages to learn the optimal policy by balancing the pole for at least 5 minutes. However, a slower decay of the exploration rate is required in order to ensure convergence for all repeats.

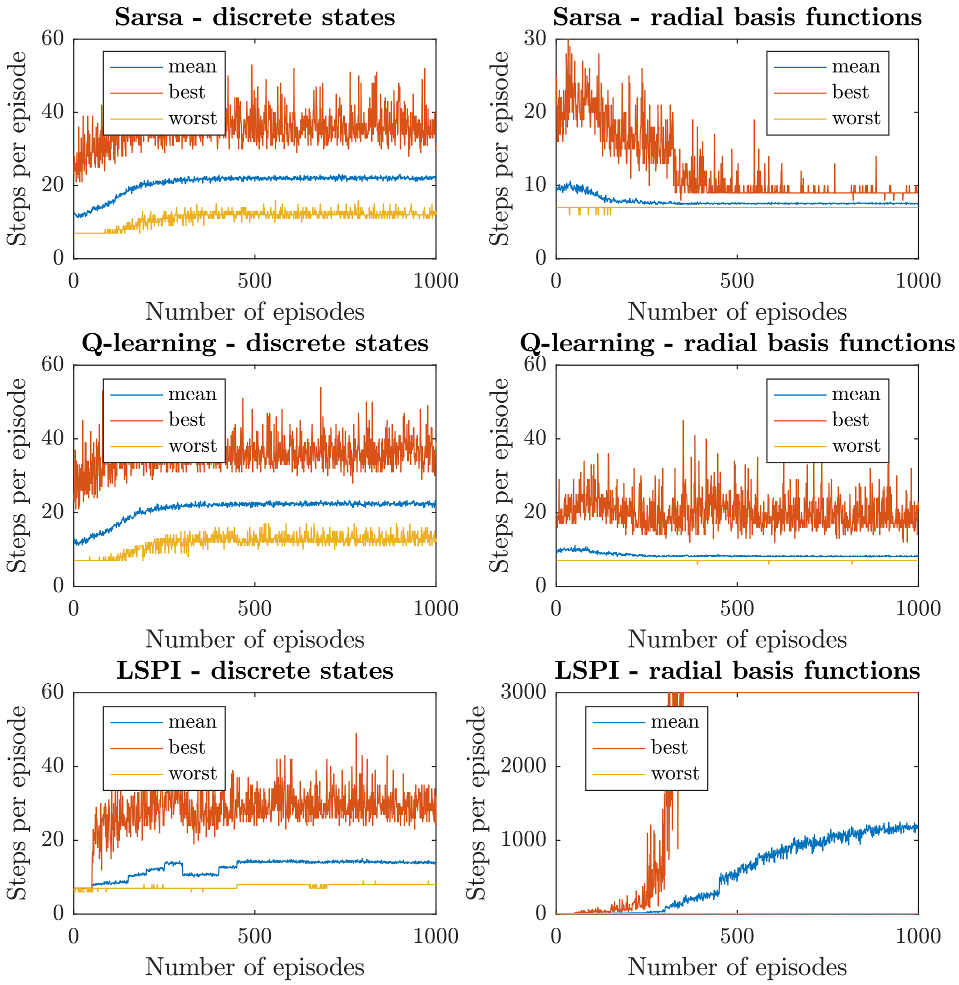


Figure 4.13: Plots of the number of steps required to reach the optimal point versus number of episodes for the three considered reinforcement learning algorithms with and without function approximation for the control of the inverted pendulum.

- While function approximation was not required for the simple grid world navigation problem, here RBFs are fundamental in decreasing the learning time due to the continuous nature of the state space.

4.7 Summary and discussion

In this chapter, the theory of reinforcement learning has been introduced. Firstly, Markov decision processes are described so as to explain the reasoning behind the development of reinforcement learning strategies. Similarly, the problem of exploration versus exploitation typical of reinforcement learning applications has been discussed. Subsequently, two different reinforcement learning methodologies have been described in detail: Monte-Carlo and temporal difference methods. Algorithmic implementations have also been presented, including four popular temporal difference schemes. In addition, function approximation has been discussed. Finally, two benchmark problems have been analysed in order to assess the performance of the Q-learning, Sarsa and LSPI algorithms, which have been mainly used during the project. Furthermore, the benchmark problems have also shown that whereas function approximation is not necessary if the problem presents a small number of discrete states, the technique is necessary to ensure learning in a feasibly small amount of episodes for larger, continuous state spaces.

In Chapter 6, reinforcement learning will be applied to the control of WECs. Firstly, a basic implementation of Monte-Carlo methods is applied to the declutching control of a point absorber. Then, Q-learning, Sarsa and LSPI algorithms are implemented for the resistive and reactive control of WECs. In particular, the time averaged control problem is expressed in a formulation similar to the grid world problem to speed up learning and aid convergence. NFQ has been applied to the control of a tidal turbine in a collaboration with other researchers at the University of Edinburgh. Since this topic goes beyond the focus of this project, it will no longer be treated within this thesis.

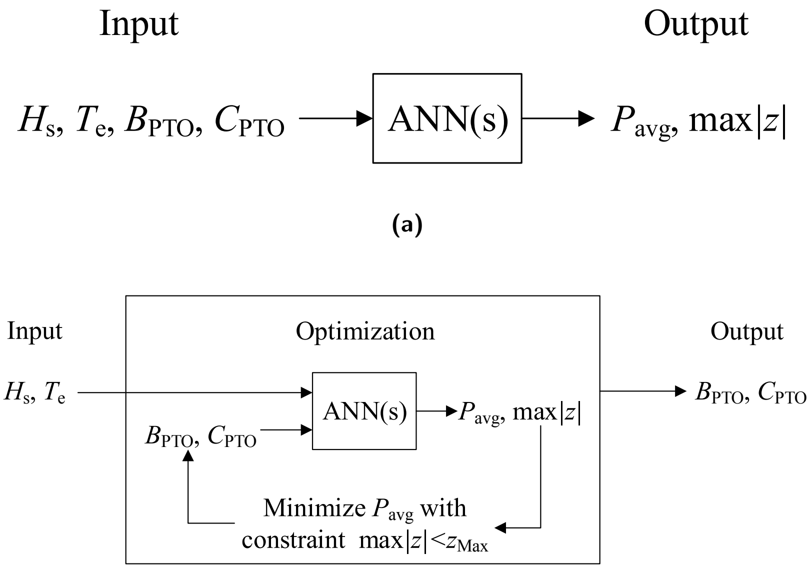
Chapter 5

Application of ANNs and optimization to the determination of PTO parameters

In this chapter, neural networks are applied to the reactive, or impedance-matching, control of WECs, which is introduced in Section 3.3. ANNs are a class of supervised learning algorithms. Fore more information, the user is referred to Appendix A. ANNs have been previously employed for the real-time system identification of WECs, for instance in the works of Giorgi *et al.* (2016b) and Valério *et al.* (2008). In particular, autoregressive ANNs with exogenous inputs and locally recurrent ANNs have been adopted. Here, a different approach is proposed.

Although ANNs are still used for system identification, the analysed system presents much slower variation. Instead of analysing the system in real time, a time-averaged, non-linear model is obtained, which maps the significant wave height, H_s, energy wave period, T_e, and PTO parameters, namely the PTO damping and stiffness coefficients (B_PTO and C_PTO, respectively) to the mean generated power, P_avg, and maximum displacement magnitude, $\max|z|$. This formulation enables the selection of the optimal PTO coefficients that maximize energy absorption while meeting displacement constraints at the start of each time interval. The length of the time interval is selected to be long enough to ensure the full decay of transient effects associated with a change in PTO coefficients. As a result, the coefficients from previous time intervals do not greatly affect the data in the current interval so that simpler feedforward ANNs can be used instead of autoregressive and local recurrent ANNs.

Hence, in order to train the ANNs, values of H_s, T_e, B_PTO, C_PTO, the mean absorbed power, P_avg, and $\max|z|$ are collected for each time horizon throughout the operation of the device, as entries of the training data set. Wave prediction techniques (for the determination of H_s and T_e) would be used in a realistic scenario to aid the selection of suitable control parameters. The estimates for P_avg and $\max|z|$ can be expressed through the functions $f\left(H_\mathrm{s}, T_\mathrm{e}, B_\mathrm{PTO}, C_\mathrm{PTO}\right)$ and $g\left(H_\mathrm{s}, T_\mathrm{e}, B_\mathrm{PTO}, C_\mathrm{PTO}\right)$ respectively.



(a)

(b)

Figure 5.1: Diagram of the ANN(s) used for the system identification of the WEC power absorption and maximum displacement (a) and diagram of the optimization strategy that relies on the ANN(s) (b).

The trained ANNs will then be fed to optimization functions in order to find the optimal PTO damping and stiffness coefficients for every new time horizon based on the forecast sea state conditions. A brute-force approach that relies on parallel processing is proposed for the optimization due to the non-linear nature of the model provided by the ANN.

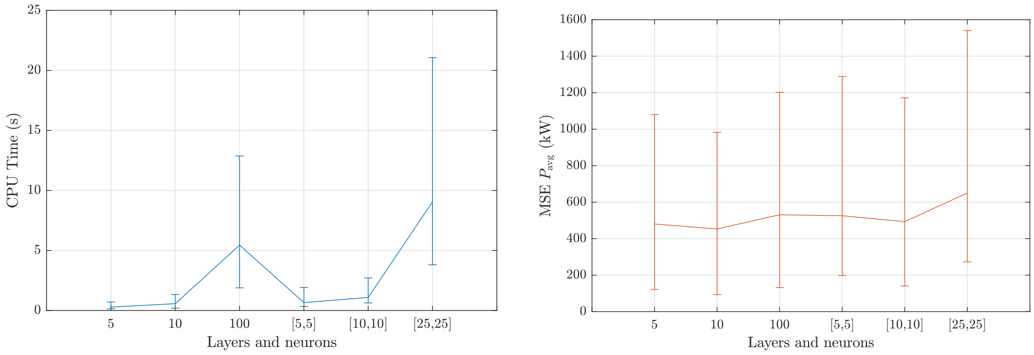
Figure 5.1a shows the proposed ANN(s) graphically, while Figure 5.1b displays the adopted optimization procedure.

The following sections will describe the proposed approach in detail. A case study, based on the simple point absorber model introduced in Section 2.3.1 is also provided to demonstrate the performance of the scheme.

5.1 Application of ANNs to the reactive control of WECs

As aforementioned, in this work ANNs are employed in order to map the mean generated power and the maximum displacement at the PTO to H_s, T_e, B_PTO and C_PTO. This is achieved through a multi-layer, feed-forward ANN with two output variables: P_avg and $\max|z|$.

In order to select a suitable size for the ANN, a preliminary study was conducted



(a) **(b)**

Figure 5.2: Mean central processing unit (CPU) time (a) and mean square error (MSE) (b) associated with the prediction of the mean generated power for different ANN configurations in terms of hidden layers and neurons for 5 weight initializations and 5 training and test sets. The upper bar corresponds to the sum of the mean value and half the standard deviation, while the lower bar to the minimum value of all cases in order to prevent negative values.

to assess the performance of possible network configurations in estimating the mean absorbed power (hence, ignoring max $|z|$ and reducing the number of output variables to one). This study has been performed with the linear model of the point absorber limited to motions in the heave degree of freedom, which was first introduced in Section 2.3.1. In particular, a single hidden layer with 5, 10, and 100 neurons, and two hidden layers with 5, 10 and 25 neurons each have been considered. For each configuration, 25 cases have been generated as the combination of 5 different random initializations of the weight matrices and 5 training and test datasets. In fact, a single training dataset has been sampled from simulations in irregular waves for the sea states in Table 5.1, which has also been used to pre-initialize the ANN-based control in Section 5.3.3. According to standard practice with ANN training (Hagan *et al.*, 1996), the whole set has been subdivided into the five distinct training and test sets by randomly reordering it, and each time selecting the first 250 points for the test set (about 10%) and the remaining 2239 samples for the training set (approximately 90%). For each case, the ANN has been trained using the training samples, and then used to estimate $P_{\rm avg}$ for the test set. The mean square error between the prediction and the actual mean generated power value has been calculated, as well as the overall computing time required for the ANN implementation described below. Afterwards, the mean and standard deviation of these values have been computed for each network configuration, and plotted in Figure 5.2.

From Figure 5.2, it is clear that the decision on the size of the ANN should be based on a compromise between performance and accuracy. On the one hand, denser networks result in greater memory requirements and computational cost, as shown in Figure 5.2a. In particular, it is interesting to notice that the configuration with two hidden layers with 10 neurons each, which contains a total of 100 connections between the two hidden layers, presents a much lower computational cost than a single layer with 100 neurons, mainly due to implementation reasons. On the other hand, the deeper the network, the greater the number of features that can be matched from the original function; similarly, the greater the number of neurons, the more complex the fitted function shape (LeCun *et al.*, 2015). An example is the lower mean square error associated with the configurations with 10 neurons as compared with those with 5 in Figure 5.2b. Nevertheless, an excessive number of neurons can result in overfitting the input data (Hagan *et al.*, 1996), i.e. fitting the random noise in addition to the underlying relationship, which is highly undesirable since the ANN is expected to generalise the shape of the P_{avg} and $\max|z|$ curves. In Figure 5.2b, this evidently occurs for a single hidden layer with 100 neurons and two hidden layers with 25 neurons each. Although a single hidden layer seems to perform best, this preliminary study has been carried out on a relatively small dataset, considering only a limited number of sea states. Therefore, it has been preferred to use a configuration with two hidden layers each with 10 neurons in order to represent the possible extra features associated with the additional sea states. Additionally, this results in only a minor increase in computational time. Similar results are obtained from the mapping of the maximum displacement.

A schematic diagram of the feed-forward ANN can be seen in Figure 5.3. The network presents an input layer with 4 neurons (one for each input variable), two hidden layers with $m = 10$ and $n = 10$ neurons each, and an output layer with two output variables. Furthermore, it is possible to see that the input and hidden layers have an additional bias term, which is required to find the intercept of the fitted functions at each stage in the ANN Hagan *et al.* (1996). Each layer l presents input and output vectors $\boldsymbol{o}^{l-1}$ and $\boldsymbol{o}^{l}$, respectively, with the input and output to the network being denoted by the vectors (or matrix, for multiple samples) $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. The signal between each two matrices is multiplied by weight matrices $\boldsymbol{W}^{l}$, with $l = 1, 2, 3$. The weight matrices for the bias terms are represented as $\boldsymbol{b}^{l}$.

For a given sample, the ANN provides a predicted output using forward propagation, as described in Section A.2.3. As shown in Figure 5.3, the hidden layers present the hyperbolic tangent activation function, while the output layer a linear function. The network is updated at regular intervals by employing a number of samples with the Levenberg-Marquard batch method devised by Hagan and Menhaj (1994), which is described in Section A.2.4.2.

It is important to notice that the input variables, i.e. H_{s}, T_{e}, B_{PTO} and C_{PTO}, need to be normalized through their mean and standard deviation before being fed to the ANNs for training. Furthermore, the mean power values have also been normalized with respect to the maximum (for positive values) and minimum (for negative values).

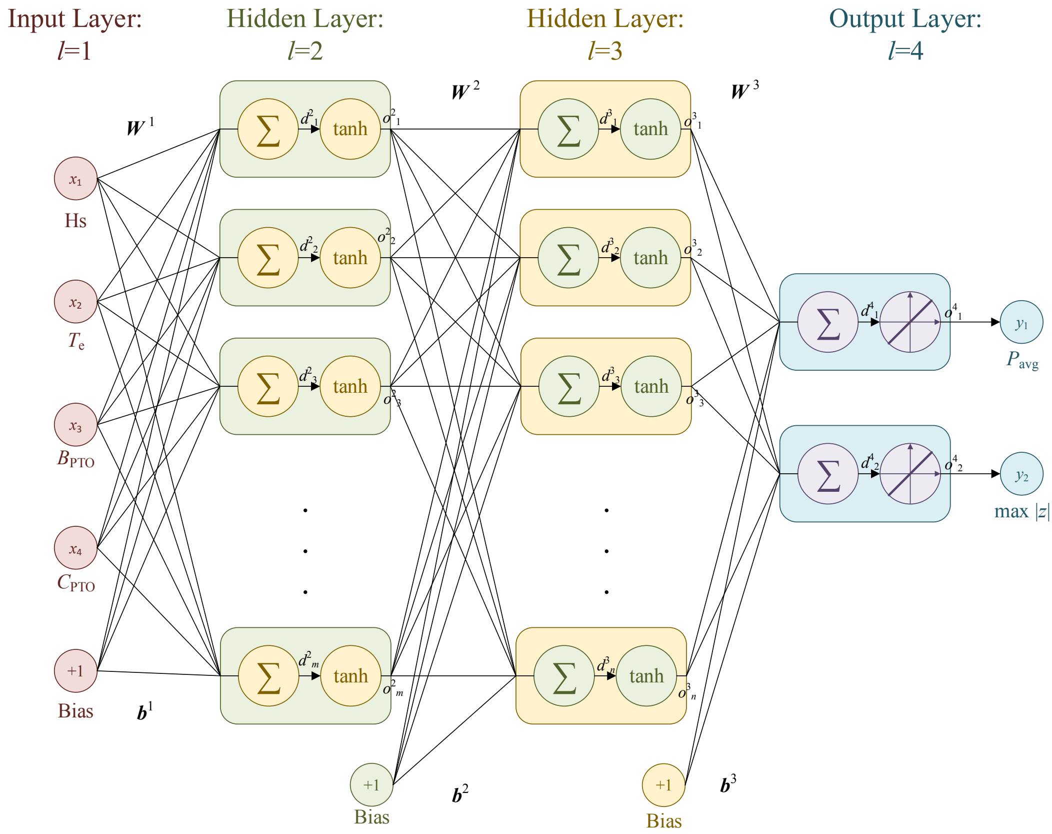


Figure 5.3: Schematic diagram of the feed-forward ANN for the approximation of the mean generated power or maximum PTO displacement.

This has been necessary because the points lying on the $B_{\mathrm{PTO}} = 0$ boundary of the search space presented excessively high negative power values that seriously affected the quality of the mapping.

5.2 Multistart optimization

At the start of every new time horizon, the controller should select the PTO damping and stiffness coefficients that will result in maximum energy extraction for the predicted sea state during the horizon, in compliance with the constraint on the PTO displacement. This is clearly a non-linear optimization problem, since both P_{avg} and $\max|z|$ are non-linear functions of H_{s}, T_{e}, B_{PTO} and C_{PTO}. In addition, the values of the PTO damping and stiffness coefficients must be bounded within sensible values, so that the problem is constrained as well.

By removing the dependence on the significant wave height and wave energy period from functions f and g due to space limitations for display purposes, the cost function can be expressed at the start of each new time horizon h as follows:

$$c(h) = \begin{cases} -f\left(B_{\mathrm{PTO}}, C_{\mathrm{PTO}}\right) & \text{if } \left|g\left(B_{\mathrm{PTO}}, C_{\mathrm{PTO}}\right)\right| \leq z_{\mathrm{Max}} \quad (5.1\mathrm{a}) \\ +1 & \text{if } \left|g\left(B_{\mathrm{PTO}}, C_{\mathrm{PTO}}\right)\right| > z_{\mathrm{Max}} \quad (5.1\mathrm{b}) \end{cases}$$

subject to:

$$B_{\min} \leq B_{\mathrm{PTO}} \leq B_{\mathrm{Max}}, \tag{5.2a}$$

$$C_{\min} \leq C_{\mathrm{PTO}} \leq C_{\mathrm{Max}}. \tag{5.2b}$$

The values of the maximum and minimum allowable PTO damping and stiffness coefficients can be derived through simulations with accurate, non-linear models during the design stage in order to prevent damage to the generator in the most energetic sea states likely to be encountered, where the WEC velocity and displacement are highest. Note that to prevent unstable behaviour of the WEC, $B_{\min} > 0$ should be selected.

Due to the non-linear nature of the mapping provided by the ANN, a global optimization scheme is necessary. Genetic and other nature-inspired algorithms have been extensively used recently for the solution of non-linear optimization problems that present multiple minima (Arora, 2012). Nevertheless, in this work, a strong emphasis is given to performance, since the optimization needs to be repeated at the start of each new time interval. For this reason, it has been preferred to use the Multistart algorithm, developed by Ugray *et al.* (2006).

This algorithm is a type of "brute-force" strategy. The technique consists in generating a number of start points, sampled randomly within the $B_{\mathrm{PTO}}, C_{\mathrm{PTO}}$ search space. From each point, a fast convex optimization scheme will be run, which will converge towards the nearest local optimum. Although convergence is not assured, a large number of starting points greatly increases the chances (Ugray *et al.*, 2006). The main advantage of this technique over alternative methods, such as global search, is its simple parallel implementation, which can result in large savings in computational time. However, note that the use of genetic algorithms in conjunction with parallel processing could be a valid alternative.

In this application of the Multisearch algorithm with parallel processing, a value of 100 starting points has been selected. From each point, an optimization is run using an interior point algorithm, which is described in Byrd *et al.* (2000) and Waltz *et al.* (2006). In particular, the Mathworks functions *MultiStart* and *fmincon* (relying on the interior point optimization algorithm (Arora, 2012)) have been used respectively. With these implementations, one Multistart optimization using the cost function in (5.1) takes 8.62 s on a quad-core, i7 computer with 16GB RAM, whereas a global search takes 29.20 s. A greater number of cores and an implementation in a lower-order language, such as C or Fortran, can result in even greater computational savings.

5.2.1 Algorithm

Figure 5.4 shows the algorithm for the ANN-based reactive control of the point absorber described in this article. As aforementioned, a time-averaged approach is used, where new values of B_{PTO} and C_{PTO} are selected at the start of every new time horizon h and applied throughout its duration $D(h)$. On the one hand, a longer duration is preferable for the power averaging and sea state statistical analysis so as to produce less noisy training data. On the other hand, a shorter time span can result in faster training. Furthermore, the controller would be able to track changes in the sea state on a smaller time scale, thus moving towards real-time control and possibly higher energy extraction. For these reasons, $D(h) = 20T_{\mathrm{e}}(h)$ has been chosen in both regular and irregular waves.

As can be seen from Figure 5.4, the first step in every time horizon is to predict the significant wave height and energy wave period during the time interval. Different approaches have been proposed for this problem, with example methods being Kalman filters, deterministic sea wave prediction (Li *et al.*, 2012), autoregressive models (Fusco and Ringwood, 2010b), and even ANNs (Shoori J. *et al.*, 2015). Although these studies analyse the wave elevation, which is forecast with accuracy only 15 s into the future (Fusco and Ringwood, 2010a), it is assumed that similar strategies can be found for the forecast of the statistical wave conditions for one time horizon. For simplicity, in this initial work the actual values for H_{s} and T_{e} have been used, since the wave traces employed in the simulations are known in advance. $H_{\mathrm{s}}(h)$ and $T_{\mathrm{e}}(h)$ are then used to

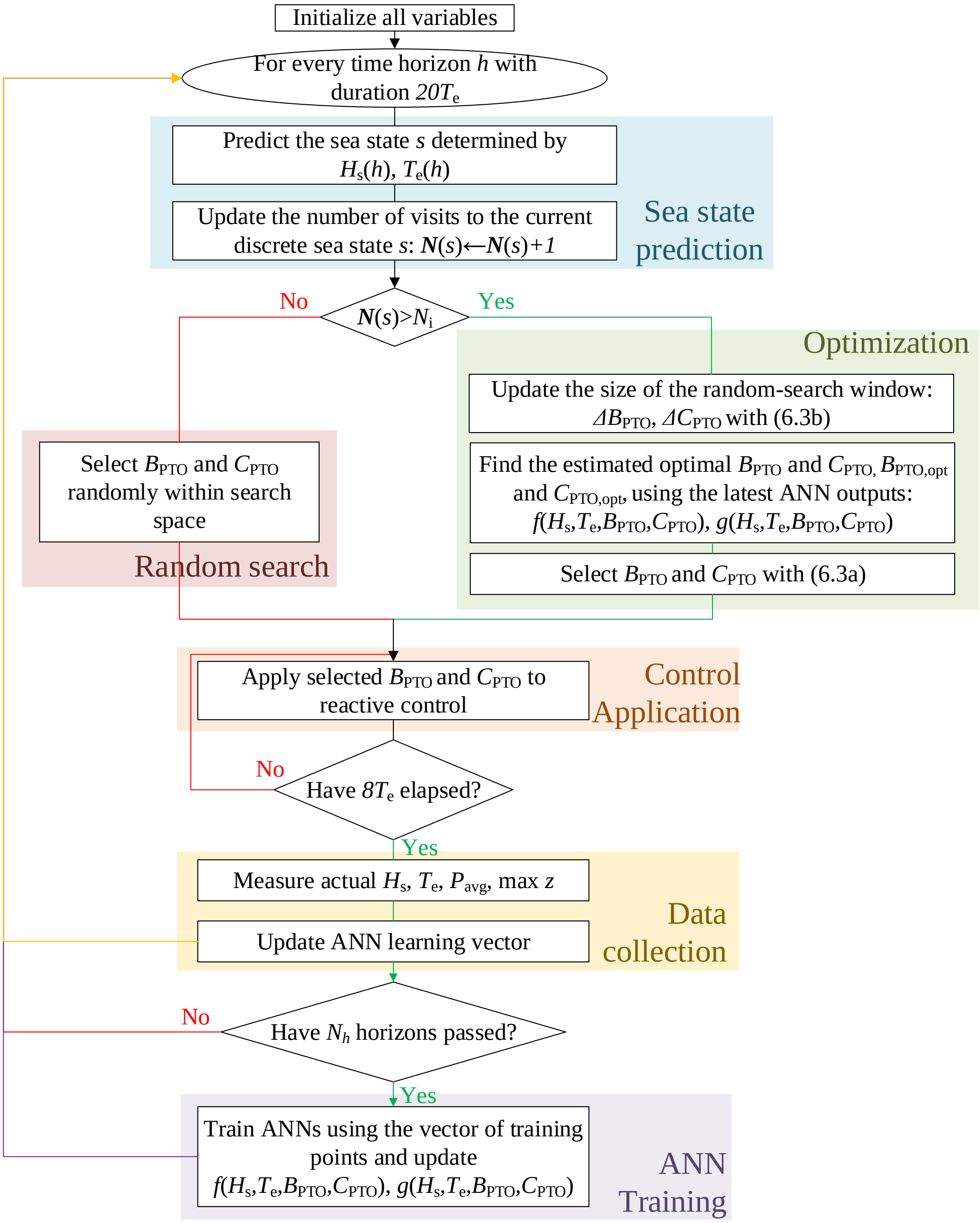

Initialize all variables
For every time horizon h with duration $20T_e$
Predict the sea state s determined by $H_s(h)$, $T_e(h)$
Update the number of visits to the current discrete sea state s: $\boldsymbol{N}(s)\leftarrow\boldsymbol{N}(s)+1$
Sea state prediction
No
$\boldsymbol{N}(s)>N_i$
Yes
Optimization
Update the size of the random-search window: ΔB_{PTO}, ΔC_{PTO} with (6.3b)
Find the estimated optimal B_{PTO} and C_{PTO}, $B_{PTO,opt}$ and $C_{PTO,opt}$, using the latest ANN outputs: $f(H_s,T_e,B_{PTO},C_{PTO})$, $g(H_s,T_e,B_{PTO},C_{PTO})$
Select B_{PTO} and C_{PTO} with (6.3a)
Select B_{PTO} and C_{PTO} randomly within search space
Random search
Apply selected B_{PTO} and C_{PTO} to reactive control
Control Application
No
Have $8T_e$ elapsed?
Yes
Measure actual H_s, T_e, P_{avg}, max z
Update ANN learning vector
Data collection
No
Have N_h horizons passed?
Yes
Train ANNs using the vector of training points and update $f(H_s,T_e,B_{PTO},C_{PTO})$, $g(H_s,T_e,B_{PTO},C_{PTO})$
ANN Training

Figure 5.4: Flow chart of the ANN-based reactive control of a point absorber.

update the count of the number of observations in the current discrete sea state, s. For this purpose a table, $\boldsymbol{N}$, is employed, with an entry for each discrete sea state with ranges of 1 m and 1 s for each dimension respectively.

During the first N_i visits to each discrete sea state, the values of the PTO damping and stiffness coefficients are selected randomly to ensure initial exploration. Once $\boldsymbol{N}(s) > N_\mathrm{i}$, the Multistart optimization can be run using the cost function in (5.1) in order to find the optimal coefficients, $B_\mathrm{PTO,opt}$ and $C_\mathrm{PTO,opt}$, for the forecast significant wave height and energy wave period. However, the ANN estimates f and g can be very inaccurate initially. For this reason, B_PTO and C_PTO are in fact selected randomly within a region around the optimum that shrinks with the number of data points collected in the sea state:

$$B_\mathrm{PTO} = B_\mathrm{PTO,opt} + \Delta B_\mathrm{PTO}, \quad \text{where} \tag{5.3a}$$

$$\Delta B_\mathrm{PTO} = (r - 0.5) \cdot \mathrm{range}(B_\mathrm{PTO}) \cdot 0.9^{\boldsymbol{N}(s)-N_\mathrm{i}}, \tag{5.3b}$$

with $r = [0, 1]$ being a random number. The same applies to C_PTO. Upper and lower bounds are used to ensure the chosen values lie within the desired range. As more data points are collected in the optimal region, the accuracy of the ANN fit will increase.

Once B_PTO and C_PTO are chosen and applied, measurements are employed to compute the mean absorbed power, maximum PTO displacement and actual $H_\mathrm{s}(h)$ and $T_\mathrm{e}(h)$ during the time interval. These values are in fact calculated using the data only after an initial time of $8T_\mathrm{e}(h)$ within the current horizon h in order to exclude the initial transient effects associated with a change in PTO parameters. This relatively long time also ensures that the time required for the Multistart optimization does not become an issue. Once the desired values are obtained, they are stored in memory as a data sample so that they can be used for training the ANN.

The ANN is trained every $N_h = 20$ time horizons, employing 90% of training points. The remaining 10% of the samples is used for validation and hence to check the quality of the fit. Each sample presents H_s, T_e, B_PTO, C_PTO as input, and P_avg and $\max|z|$ as output. The larger the number of training points, the less the risk of overfitting the data and the more accurate the estimates of the ANN. However, this will also cause an increase in training time and, more importantly, it may result in an excessive memory requirement. Therefore, for a practical application, it is expected that the number of training points will be limited to a large number, say 10^6. Care will be needed in order to ensure that a similar number of data points is kept for each discrete sea state when overriding old data with new readings, as well as to explore a broad range of B_PTO and C_PTO values so as to aid the training of the ANN.

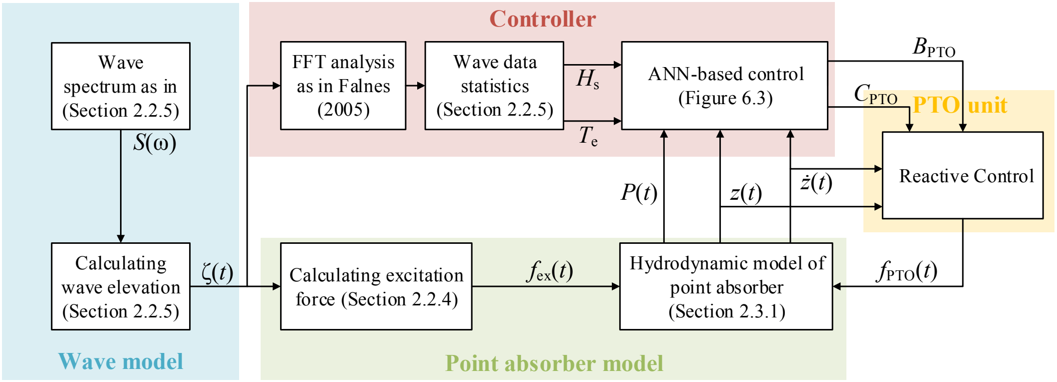

Controller
Wave spectrum as in (Section 2.2.5)
$S(\omega)$
FFT analysis as in Falnes (2005)
Wave data statistics (Section 2.2.5)
H_s
T_e
ANN-based control (Figure 6.3)
B_{PTO}
C_{PTO}
PTO unit
Reactive Control
$P(t)$
$z(t)$
$\dot{z}(t)$
Calculating wave elevation (Section 2.2.5)
$\zeta(t)$
Calculating excitation force (Section 2.2.4)
$f_{ex}(t)$
Hydrodynamic model of point absorber (Section 2.3.1)
$f_{PTO}(t)$
Wave model
Point absorber model

Figure 5.5: Flow chart of the program used in the simulations of the point absorber with ANN reactive control.

5.3 Simulation Results

5.3.1 Simulation system

The performance of the proposed algorithm has been assessed with simulations of the simple point absorbers restricted to motions in heave, introduced in Section 2.3.1. The PTO force saturation and float displacement limits have been set to 1 MN and ±5 m, respectively. A PTO efficiency of 75% is assumed.

Figure 5.5 shows graphically the the program used for the simulation of the WEC. Instead of sensors installed on a wave buoy, in the simulations a wave model provides the recorded wave elevation.

The search space has been limited to within $B_{\text{min}} = 0$ and $B_{\text{Max}} = 2$ MNs/m, and $C_{\text{min}} = -1$ MN/m and $C_{\text{Max}} = 0$ for the PTO damping and stiffness coefficients, respectively. A wider search space has been selected for the PTO damping coefficient in order to prevent damage in large waves, when greater damping and no stiffness are required as shown in Section 3.5. Nevertheless, the larger the search space, the longer the learning time; hence, an excessive search space needs to be avoided.

For the first 15 minutes of the simulations, no control force is applied in order to let the system dynamics settle. For this reason, all wave traces are in fact generated with an extra 15-minute interval at the start.

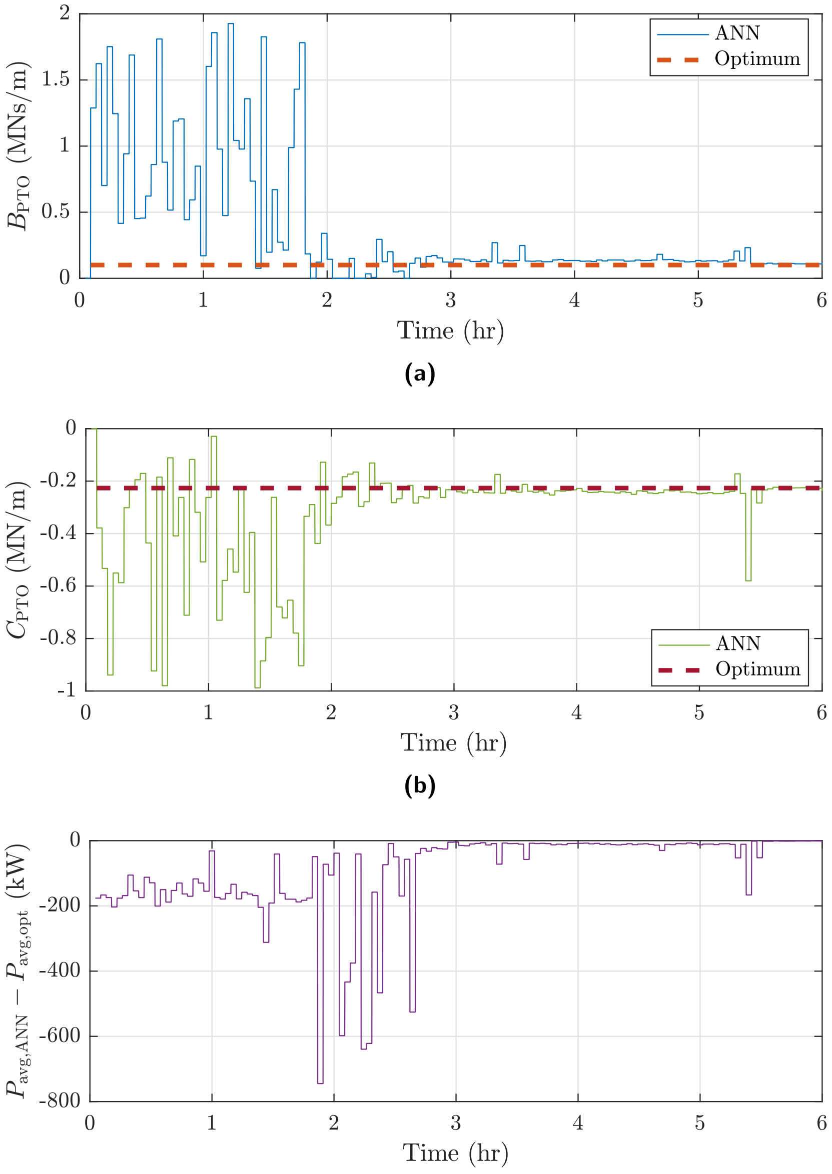
5.3.2 Results in regular waves

In regular waves, a 6-hour-long wave trace with unit amplitude and a wave period of 8 s has been analysed. As can be seen in Figures 5.6a and 5.6b, the ANN-based algorithm learns successfully the optimal PTO damping and stiffness coefficients respectively. The optimal values (dotted lines) have been obtained by running a Nelder-Mead optimization with the time domain model described in Section 2.3.1 and Chapter 3 (hence, the optimization is run with the WEC model, without relying on ANNs). In particular, a wave trace lasting 20 minutes has been considered. It should be noted that the time-domain rather than the frequency-domain model has been used for the computation of the optimal PTO coefficients and corresponding absorbed power so that the force constraint could be incorporated. Figure 5.6c shows the difference in the mean power generated with ANN-based control and state-of-the-art reactive control, where $P_{\mathrm{avg,opt}} = 176.24$ kW. A value of $N_{\mathrm{i}} = 40$ has been used.

5.3.3 Results in irregular waves

In irregular waves, even within a single sea state, the significant wave height and wave energy period do vary, if they are measured within a short time interval like $20T_{\mathrm{e}}$. On the one hand, it would be nice to show convergence of the algorithm in one sea state, and then build onto learning in multiple sea states. On the other hand, the accuracy of ANNs is greatly improved and the effects of overfitting greatly reduced the wider the range of their samples (Hagan *et al.*, 1996) and thus the wider the range of sea conditions. For this reason, the proposed ANN-based reactive control algorithm is run for the 9 wave traces shown in Table 5.1. Each wave time series is generated with a Bretschneider spectrum (thus, broad-banded) (Holthuijsen, 2007) and lasts 3 hours. Although these wave traces have been simulated independently due to computational constraints, they should be treated as a continuous time series where 9 independent sea states are observed in the order provide in Table 5.1, with a value of $N_{\mathrm{i}} = 120$ being used. In particular, for each wave trace the list of samples is initialized with the values observed in the previous runs. The series of a sea states is repeated another time but with a different seed number to the random number generator for a total wave trace with an overall duration of 54 hours (excluding the 15 minutes required for the initialization of each wave trace).

The learning behaviour of the proposed ANN-based reactive control algorithm in irregular waves is displayed in a compact way in Figure 5.7. The figure shows the controller performance for the first wave trace, i.e. $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 8$ s. In particular, the very first run (when the list of samples is empty at the start) is shown with dotted lines and labelled as "initial", since learning has just been initialized. The system is simulated in the same wave conditions again *after* the control has been applied for 54 hours in



(a)

(b)

(c)

Figure 5.6: PTO damping (a) and stiffness (b) coefficients obtained from the ANN-based control as compared with the optimal value in regular waves with $H_s = 2$ m and $T_e = 8$ s. (c) shows the difference in the corresponding mean generated power.

Table 5.1: Significant wave height, energy wave period and duration of the wave traces used for the analysis of the ANN-based control in irregular waves.

H_s (m)	2	1	1	1	2	2	3	3	3
T_e (s)	8	8	9	10	9	10	10	9	8
duration (hr)	3	3	3	3	3	3	3	3	3
no. repetitions	2	2	2	2	2	2	2	2	2

the wave traces shown in Table 5.1. The corresponding performance is shown with continuous lines in Figure 5.7 and labelled as ”trained”, since learning has completed by then with a large number of samples being available for the training of the ANN. Furthermore, in this case the exploration rate has almost fully decayed, as the discrete sea state has already been experienced for 6 hours. Additionally, the optimal value for the PTO coefficients and the corresponding absorbed energy is calculated running a MultiStart optimization of the WEC model in the same wave trace.

5.4 Discussion

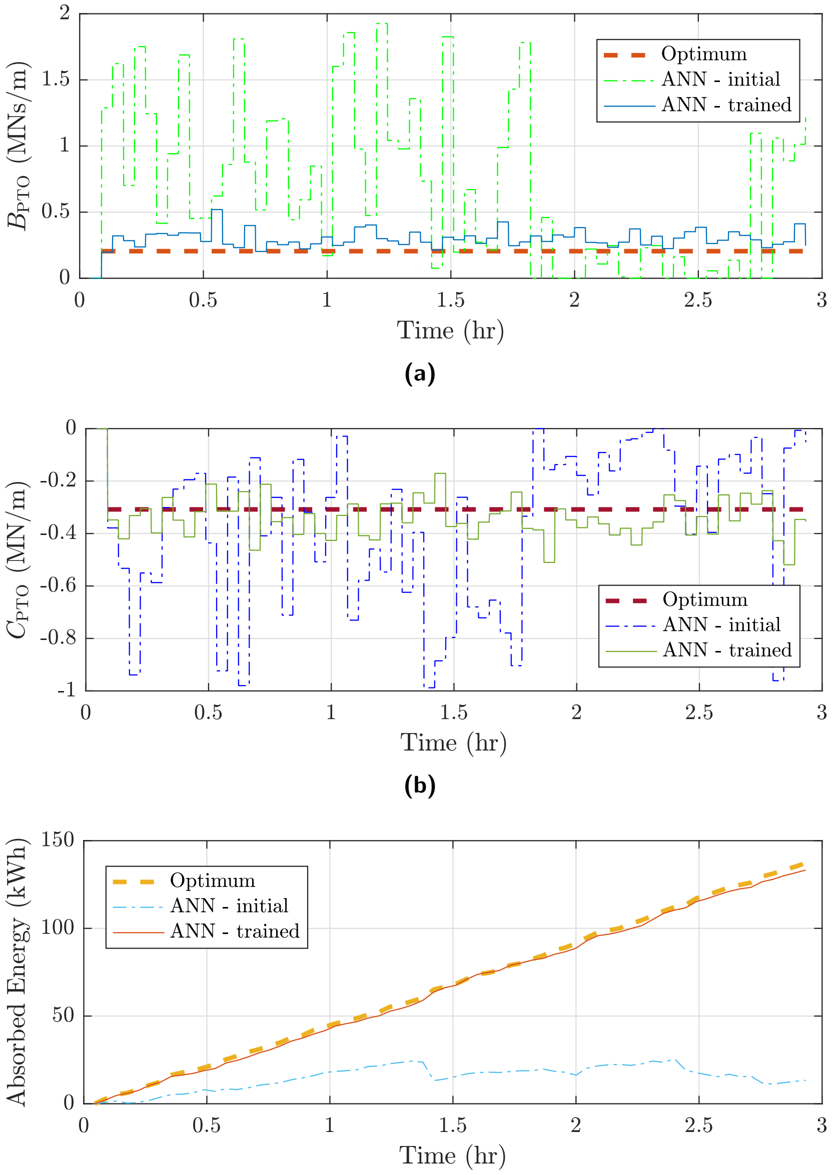
5.4.1 Regular waves

As shown in Figure 5.6, the ANN-based algorithm learns the optimal PTO damping and stiffness coefficients in regular waves within 4 hours after being randomly initialized. In the figures, it is possible to recognize three distinct regions: an initial region where completely random actions are selected ($\boldsymbol{N}(s) \leq N_i$), a section where random actions are taken around the expected optimum within a shrinking range (until $0.9^{\boldsymbol{N}(s)-N_i} \rightarrow 0$), and a final part where convergence has been reached. Within this last region, it is interesting to notice three random points (after approximately 5.5 hours). These are caused by the Multistart algorithm converging towards the wrong local optimum in the corresponding time horizons. This is a possibility that needs to be taken into account when designing the control for an actual device, with its probability decreasing with the number of starting points. Nevertheless, the low computational cost means this optimization method is still preferred over global search or genetic algorithms. Oddly, the three random points also provide the ANNs with the missing training points for perfect convergence to the optimal PTO coefficients.

5.4.2 Irregular waves

The convergence of the algorithm to the optimal PTO coefficients in irregular waves is shown by the ”trained” lines in Figure 5.7. Oscillations in the values obtained with the ANN-based control are due to changes in wave conditions over the smaller time scale of $20T_\mathrm{e}$. The energy absorption is almost identical to state-of-the-art reactive control applied using the optimal coefficients for the WEC model in this wave trace.

At first sight, 54 hours may seem like a very long learning time. However, this corresponds to 6 hours of learning time for each sea state, which is much more realistic. Once a sufficient number of points is obtained, the ANN can generalise the information to unseen sea states, thus further reducing the overall learning time. In addition, the



(a)

(b)

(c)

Figure 5.7: PTO damping (a) and stiffness (b) coefficients adopted by the ANN-based control at the start and after 54-hours of training in the wave conditions shown in Table 5.1 in irregular waves with $H_{\text{s}} = 2$ m and $T_{\text{e}} = 8$ s. Additionally, (a) and (b) display the optimal coefficients computed from a time-domain simulation in a 20-minute-long wave trace. The corresponding curves for the absorbed energy are plotted in (c).

convergence time should be assessed in the context of the lifetime of a WEC, which is expected to be 20 to 25 years long (Cruz, 2008).

In this work, discrete sea states have been analysed, each lasting 3 hours due to practical issues with the code implementation. In reality, the energy content in waves changes uniformly in time (hence, not through discrete sea states), with the duration of a typical sea state being 0.5 to 6 hours (Holthuijsen, 2007). Since P_{avg} and $\max|z|$ can be considered to be purely dependent on the values of B_{PTO}, C_{PTO}, H_{s} and T_{e} in the *current* time interval, the samples of the ANN are independent of past data. Therefore, the algorithm can be safely applied to realistic, continuously varying wave conditions. In fact, the quality of the mapping provided by the ANN is expected to improve in continuously varying sea states, which result in a broader range of samples (Hagan *et al.*, 1996). Furthermore, under realistic wave conditions, the ANN-based reactive control is expected to result in higher energy absorption than coefficient-based reactive control, since the latter uses a look-up table with discrete sea states, thus being less responsive to changes in wave energy over a shorter time scale. Additionally, the ANN-based method can adapt to changes in the device dynamics with time, e.g. due to marine growth or non-critical subsystem failure.

5.5 Chapter summary

In this chapter, an adaptive algorithm for reactive control of a WEC has been proposed, which relies on ANNs. The approach uses a time-averaged approach, where the PTO damping and stiffness coefficients are optimized for each sea state, as given by the significant wave height and energy wave period. The time-averaging interval lasts 20 wave cycles. A neural network is employed to provide a mapping between the mean generated power and maximum observed displacement amplitude, and the significant wave height, energy wave period, the PTO damping and stiffness coefficients. The ANN thus provides a non-linear model for the system in an example of system identification. A global optimization scheme is then used to find the optimal coefficients at the start of each new time interval based on the predicted wave conditions. The algorithm has been shown to learn the optimal coefficients in each sea state in less than 6 hours per sea state, with 9 different sea states being analysed.

Although the implementation of the proposed algorithm would be feasible in a real WEC, some issues need to be analysed more in detail.

- Although using a relatively long time interval ensures the convergence of the learning behaviour, energy content in waves varies with wave group. Hence, the time interval length should be adapted to the duration of different, recognizable wave groups. This information can be obtained by a network of wave buoys ahead of the device, which would also provide accurate prediction for the incoming waves. Therefore, the algorithm should be further developed to include these considerations.

- The behaviour of the algorithm under PTO force saturation and displacement constraints should be addressed in greater detail. A move towards the treatment of wave groups should result in superior treatment of constraints, but this needs to be studied thoroughly.
- At the moment, the imposition of a displacement constraint based on the maximum displacement over the averaging period leads to conservative operation and thus a loss in the absorbed energy. Solutions can be either a move towards real-time control or the investigation of less conservative approaches in the application of the displacement constraints.

In the next chapter, the application of reinforcement learning to the declutching, resistive and reactive control of a WEC is studied as an alternative to ANNs.

Chapter 6

Application of reinforcement learning to the control of wave energy converters

In this chapter, reinforcement learning is applied to the control of WECs for the first time. First of all, the declutching control of a point absorber is analysed in Section 6.1 using a simple Monte-Carlo implementation. Work on this problem was initiated by a different researcher at Pelamis Wave Power Ltd and completed by the student. Nevertheless, as described below, this approach is found to be overly simplistic. For this reason, reinforcement learning was then employed to find the optimal PTO coefficients for each sea state in resistive and reactive control, as described in Sections 6.2 and 6.3, respectively. In this application, reinforcement learning has been found to learn successfully in a reasonable time frame, showing adaptability to changing system dynamics, e.g. to non-critical subsystem failure or marine growth. Different algorithms have also been analysed to assess their convergence properties.

The structure of this chapter is summarized in Table **??**.

6.1 Monte-Carlo methods for declutching control

As described in Chapter 4, in reinforcement learning the controller learns an optimal policy from direct interactions with the environment by taking an action in a specific state and observing the reward, or return for Monte-Carlo methods. Since declutching control presents a bang-bang type of actuation (Chapter 3), identifying the action space is simple. Conversely, declutching control requires the identification of an optimal timing for the action selection in order to maximize energy absorption. Hence, the choice of suitable state variables is more complex to determine. Furthermore, it is important to remember that Monte-Carlo methods are episodic (Chapter 4), so that the real-time implementation will need to be divided into episodes. Nevertheless, their main advantage is that they have been proved to converge towards the optimal policy for an

Table 6.1: Application of reinforcement learning algorithms to the control of WECs within this chapter.

Section	RL method	WEC control type	WEC type	Comments
6.1	Monte-Carlo	declutching	linear point absorber	single sea state analysis
6.2.3	Q-learning	resistive	linear point absorber	single and multiple sea state analysis, discrete RL states
6.2.4	LSPI	resistive	point absorber with non-linear PTO	single and multiple sea state analysis, discrete RL states and function approximation, changes in system dynamics
6.2.4	Q-learning	resistive	point absorber with non-linear PTO	single sea state analysis, discrete RL states, changes in system dynamics
6.2.4	Sarsa	resistive	point absorber with non-linear PTO	single sea state analysis, discrete RL states, changes in system dynamics
6.3.7	Q-learning	reactive	point absorber with two bodies	single and multiple sea state analysis, discrete RL states
6.3.8	LSPI	reactive	linear point absorber	single and multiple sea state analysis, discrete RL states

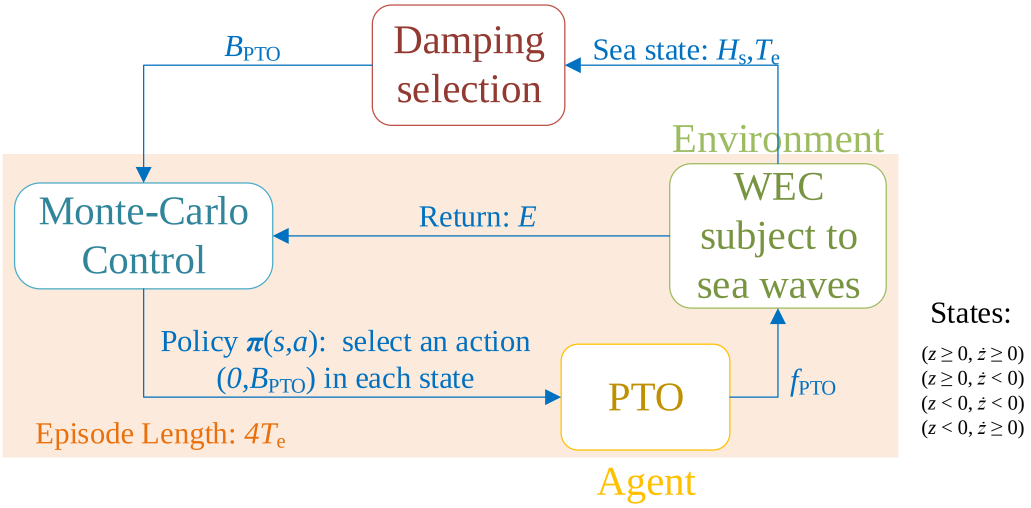

B_{PTO}
Damping selection
Sea state: H_{s},T_{e}
Environment
Monte-Carlo Control
Return: E
WEC subject to sea waves
States:
$(z \geq 0, \dot{z} \geq 0)$
$(z \geq 0, \dot{z} < 0)$
$(z < 0, \dot{z} < 0)$
$(z < 0, \dot{z} \geq 0)$
Policy $\boldsymbol{\pi}(s,a)$: select an action (0,B_{PTO}) in each state
PTO
f_{PTO}
Episode Length: $4T_{\mathrm{e}}$
Agent

Figure 6.1: Block diagram of the Monte-Carlo declutching control of the point absorber.

infinite number of visits of each state-action pair (Sutton and Barto, 1998). In the next sections, the control implementation for the point absorber introduced in Section 2.3.1 is described.

6.1.1 Formulation of Monte-Carlo declutching control

The work flow of the proposed declutching control of the point absorber is shown in Figure 6.1. The declutching control is modelled in this simple case as the actuation of either resistive control, so that the PTO force is $f_{\mathrm{PTO}} = B_{\mathrm{PTO}}\dot{z}$, with damping PTO coefficient B_{PTO}, or no control, so that $f_{\mathrm{PTO}} = 0$.

As can be seen from Figure 6.1, an episode can be defined as lasting N_T wave cycles, where a value of 4 is selected due to the randomness of irregular waves. The figure also shows the state, action and reward selection, which will be described in the following sections.

Furthermore, an external controller feeds the PTO damping coefficient to the algorithm based on the sea state measurements from a neighbouring buoy. The significant wave height, H_{s}, and energy wave period, T_{e}, are computed from a Fast Fourier Transform analysis (Holthuijsen, 2007). Hence, in fact the states, shown on the bottom right corner of the figure, have an extra dimension given by the total number of discrete states selected. This means that there may be a different optimal policy for each sea state. For simplicity, in this section, the method is presented for a single sea state. It can be then easily extended to multiple sea states by adding an extra dimension to the state-space. This will be described in greater detail in Section 6.2 and 6.3 for the application of temporal difference methods to the resistive and reactive control of WECs.

6.1.1.1 State variables

The choice of suitable reinforcement learning states is particularly important, since it will determine the timing of the selection of particular actions. In this work, a practical approach is adopted, taking into account only the variables that can be measured online with accuracy.

On the one hand, the motions of the device, derived from on-board accelerometers, are available in real time. On the other hand, the wave elevation, which would be an interesting state variable if its phase difference with the body displacement is analysed, is difficult to be observed accurately at the location of the device itself, since it is measured by a neighbouring wave buoy. The exception would be in the case of perfectly two-dimensional waves, which may occur in a channel. Therefore, the vertical displacement and velocity of the point absorber, which correspond to those at the PTO, have been taken as state variables. In particular, the states are determined by the combination of their signs, so that there are four states per wave period, or 16 states per episode. The four distinct states can be seen in the bottom left corner of Figure 6.1.

It is clear that this description of the state-space is overly simplistic, despite its possible simple, practical implementation. In general, the wave elevation (and even better the resulting wave excitation force) should be included in the state space for the maximization of energy absorption. The phase difference between the excitation force and the body velocity is of particular interest (Falnes, 2005). Indeed, the proposed implementations of declutching control with optimal command theory rely on future information on the wave elevation at the position of the device (Babarit *et al.*, 2009; Clément and Babarit, 2012). Nevertheless, here the more practical approach is selected. This issue will be described in greater detail in Section 6.1.5.

6.1.1.2 Action space

In this simple implementation of declutching control, the action space comprises of two actions, namely whether to apply no or a resistive control force, as shown in Figure 6.1. In an actual device, or in a more complex model of the hydraulic PTO, the damping term will be substituted by the operation (opening and closing) of the valves that connect the accumulators to the hydraulic circuit as well as the bypass valve.

6.1.1.3 Return function

Within this thesis, the aim of the control of a WEC is assumed to be the maximization of its energy absorption. In fact, the development of a survival strategy, the reduction of loads on the structure and the consequent increase in fatigue life are at least as, if not more, important. Nevertheless, for simplicity, at this stage no penalties are applied for large displacements, although this will be implemented in the future. Therefore, the return function should be the energy extracted over an episode i:

$$E(i) = \int_{t_\mathrm{i}(i)}^{t_\mathrm{f}(i)} P(t)\mathrm{d}t, \tag{6.1}$$

where t_i and t_f are the initial and final times of the episode. Nevertheless, due to the stochastic nature of irregular waves, there can be significant variations in the energy measured over multiple episodes for the same policy. Even within the same sea state, there are individual wave cycles with varying wave height and period, and thus wave power, as energy is transported in wave groups (Holthuijsen, 2007). Therefore, in order to ensure convergence towards the optimal policy even in irregular waves, a table $\boldsymbol{S}$ is created where the energy values are stored (i.e. summed) for each state-action combination. These correspond to all the possible policies. Since there are 2 possible actions in 4 different states, there is a total of $2^4 = 16$ possible combinations, and thus entries in the table. A separate table $\boldsymbol{N}$ is also created, which contains the total number of values per entry. Hence, the mean energy per states-actions combination can be found as follows ($\oslash$ indicates element-wise division):

$$\boldsymbol{E} = \boldsymbol{S} \oslash \boldsymbol{N}. \tag{6.2}$$

The mean energy can be very similar for a number of policies close to the optimal one. As a result, the learning time can be very long. For this reason, in order to speed up convergence, the mean energy values are first normalized and then raised to a power in order to obtain the return function for each policy. Therefore, for the current policy j the return function is given by:

$$r = \left(\frac{\boldsymbol{E}(j)}{\max_{k=1:16} \boldsymbol{E}(k)} \right)^g. \tag{6.3}$$

A value of $g = 25$ has been used in this work in order to speed up the learning time.

6.1.1.4 Exploration strategy

As discussed in Chapter 4, the policy π represents the mapping of actions to states. This means that the policy indicates which actions should be taken in which states. In Monte-Carlo methods, the policy needs to be selected at the start of each episode. It has been decided to choose a unique policy for each episode, so that the policy is repeated four times (i.e. four wave cycles) every episode. As a result, all states are visited four times in every episode.

The ϵ-greedy policy in (4.9) is selected to ensure sufficient exploration at the start of the task, with the focus shifting towards exploitation as learning progresses. This is achieved by decreasing the exploration rate after every episode as

$$\epsilon = \epsilon_0 0.9^{\boldsymbol{N}_{\mathrm{c}}(j,l)}, \tag{6.4}$$

where $\epsilon_0 = 0.5$ is the initial exploration rate, $\boldsymbol{N}_{\mathrm{c}}$ is a table that contains the number of visits to each combination of states and actions (j) and each discrete sea state (l).

Due to the small number of possible state-action combinations (16), it has been decided to ensure exploration of the complete state-action space by selecting each combination in turn for a total of four times. This results in the first 64 episodes (or $256T_{\mathrm{e}}$ in time units) presenting a fixed policy. Although this can result in a delay in the learning time, this initial forced exploration ensures the correct policy is learned.

6.1.1.5 Monte-Carlo scheme for declutching control

With the exploring-starts Monte-Carlo algorithm, it is necessary to create a three-dimensional list $\boldsymbol{R}$, which has 4 rows (number of states) and 2 columns (number of actions). At the end of each episode i, the return r is appended to $\boldsymbol{R}(s, a)$, where s, a indicate the state-action pairs that have been observed during the episode (Sutton and Barto, 1998), as given by the policy π. An every-visit Monte-Carlo scheme is used, where the return is calculated over all visits (or encounters) of the state-action pairs rather than just the first one (first-visit Monte-Carlo) (Sutton and Barto, 1998).

The state-action values are calculated at the end of each episode for each state-action pair as the mean of all returns, as shown in Algorithm 1:

$$\boldsymbol{Q}(s, a) \leftarrow \text{average}(\boldsymbol{R}(s, a)), \quad s \in \mathcal{S}, a \in \mathcal{A}. \tag{6.5}$$

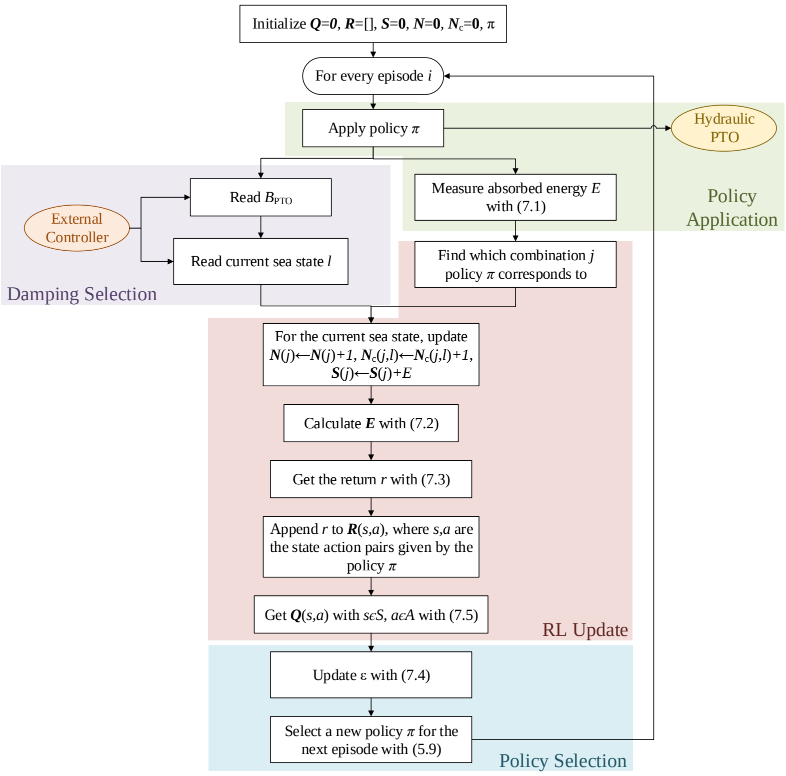

Initialize $\boldsymbol{Q}$=0, $\boldsymbol{R}$=[], $\boldsymbol{S}$=0, $\boldsymbol{N}$=0, $\boldsymbol{N}_c$=0, π
For every episode i
Apply policy π
Hydraulic PTO
Read B_{PTO}
External Controller
Read current sea state l
Damping Selection
Measure absorbed energy E with (7.1)
Policy Application
Find which combination j policy π corresponds to
For the current sea state, update $\boldsymbol{N}(j)\leftarrow\boldsymbol{N}(j)+1$, $\boldsymbol{N}_c(j,l)\leftarrow\boldsymbol{N}_c(j,l)+1$, $\boldsymbol{S}(j)\leftarrow\boldsymbol{S}(j)+E$
Calculate $\boldsymbol{E}$ with (7.2)
Get the return r with (7.3)
Append r to $\boldsymbol{R}(s,a)$, where s,a are the state action pairs given by the policy π
Get $\boldsymbol{Q}(s,a)$ with $s\epsilon S$, $a\epsilon A$ with (7.5)
RL Update
Update ε with (7.4)
Select a new policy π for the next episode with (5.9)
Policy Selection

Figure 6.2: Flowchart of the Monte-Carlo algorithm for the declutching control of the point absorber.

6.1.1.6 Algorithm

The proposed algorithm for the declutching control of the point absorber can be seen in Figure 6.2. The algorithm is a summary of the items discussed within this section. It summarises all points raised within this section. As can be seen from the figure, the PTO damping coefficient is selected for each sea state by an external controller. This value can be either obtained from simulations, as is common practice in the wave energy industry (for instance, at Pelamis Wave Power Ltd), or through learning techniques, as described in Section 6.2.

6.1.2 Simulation results

The point absorber described in Section 2.3.1 is analysed to demonstrate the learning capabilities of the proposed algorithm. A finer time step of 0.01 s is adopted due to the non-linear nature of declutching control. For simplicity, the algorithm is tested in a single sea state in both regular and irregular waves. In regular waves, a two-hour long wave trace with unit amplitude and a wave period of 8 s is generated. In irregular waves, a three-hour long wave trace with a Bretschneider spectrum with $H_\mathrm{s} = 2$ m and $T_\mathrm{p} = 9.25$ s, corresponding to $T_\mathrm{e} = 8$ s from a spectral analysis, is generated. In order

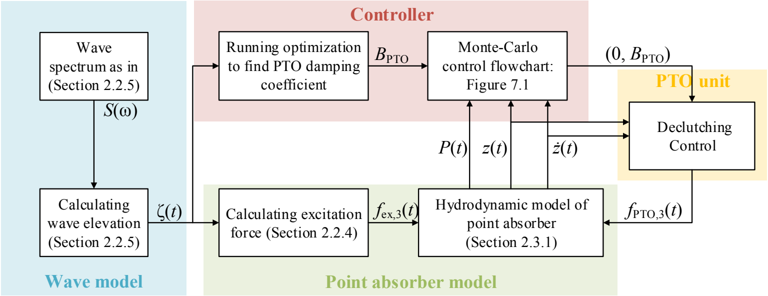

Controller
Wave spectrum as in (Section 2.2.5)
$S(\omega)$
Running optimization to find PTO damping coefficient
B_{PTO}
Monte-Carlo control flowchart: Figure 7.1
$(0, B_{\text{PTO}})$
PTO unit
Declutching Control
$P(t)$
$z(t)$
$\dot{z}(t)$
Calculating wave elevation (Section 2.2.5)
$\zeta(t)$
Calculating excitation force (Section 2.2.4)
$f_{\text{ex},3}(t)$
Hydrodynamic model of point absorber (Section 2.3.1)
$f_{\text{PTO},3}(t)$
Wave model
Point absorber model

Figure 6.3: Work-flow diagram of the program used to simulate the point absorber.

to ensure the motions of the model are fully developed, the Monte-Carlo algorithm is run only after 15 minutes from the start of the time series.

An optimization is run in order to find the optimal PTO coefficient in each sea state using the time-domain model in 20-minute-long wave traces, as described in Section 5.3. For the regular waves, a PTO damping coefficient of 305.678 kNs/m is adopted, while for the irregular wave trace a very similar PTO damping coefficient of 308.347 kNs/m is chosen. These values correspond to the optimal PTO damping coefficients in regular waves with $T = 8$ s and irregular waves with $T_{\text{e}} = 8$ s, respectively.

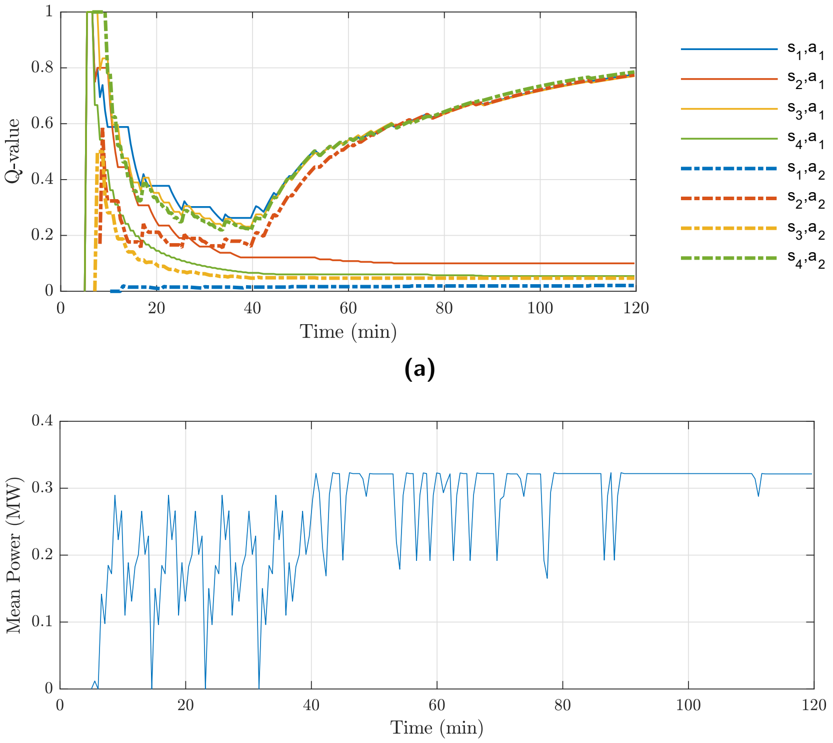
The program used to simulate the point absorber is summarised in Figure 6.3 for clarity.

6.1.3 Results in regular waves

Regular waves have been analysed first in order to assess the convergence properties of the proposed reinforcement learning control under deterministic conditions. Figure 6.4a shows the variation of the entries of the Q-table with time. In Figure 6.4b, it is possible to see the corresponding absorbed energy per episode.

The optimal policy found in Figure 6.4a once convergence is achieved is shown in Figures 6.5a and 6.5b for a short wave trace. Figure 6.5b also shows the time series of the instantaneous generated power. Figure 6.5c shows the corresponding absorbed energy as compared with optimal resistive control.

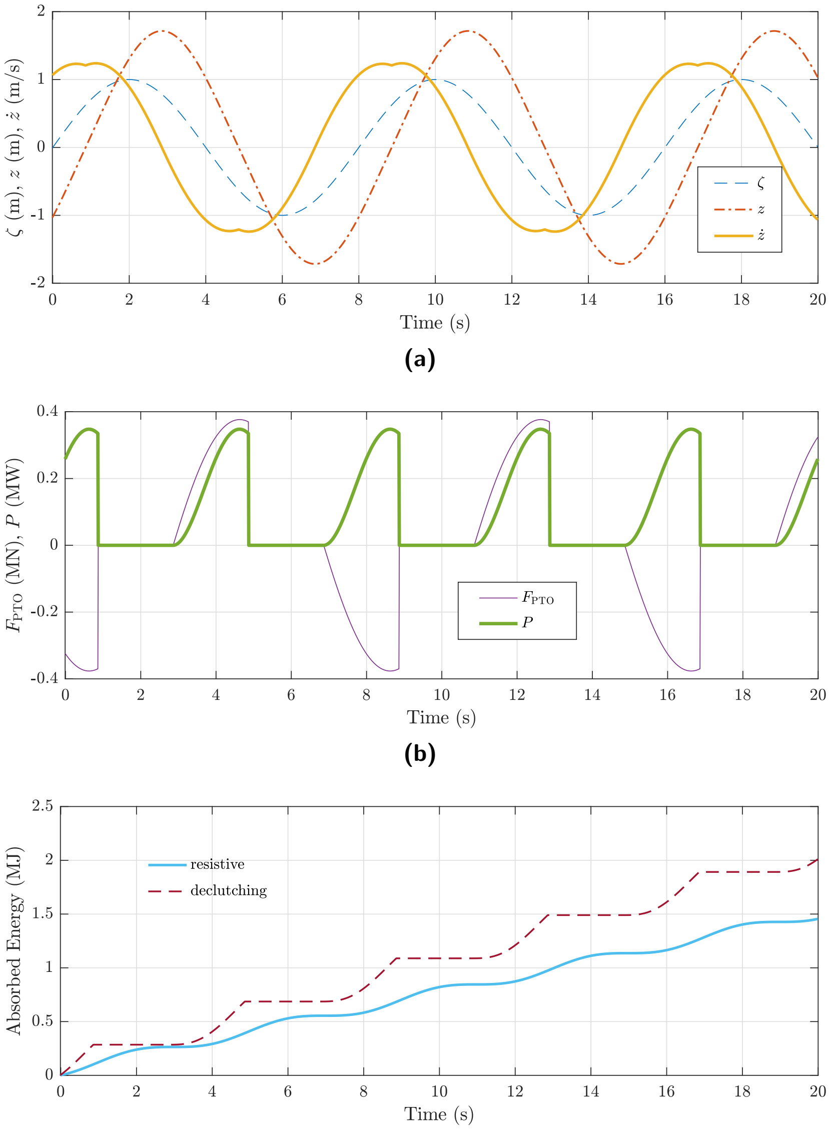
Then, a range of periods is analysed. For most wave periods, the optimal policy is unchanged: apply resistive control for $z \geq 0, \dot{z} < 0$ and $z < 0, \dot{z} \geq 0$, and no PTO force for the remaining states. However, for $T_{\text{e}} = 9, 7, 11$ s a different optimal policy is found: always apply resistive control except when $z < 0, \dot{z} \geq 0$. Since no penalty is applied for large displacements, only one wave amplitude has been considered (1 m). Figure 6.6 shows the variation in the capture width ratio of the point absorber with wave period when the optimal policy in each sea state is adopted. The performance of Monte-Carlo declutching control is compared with that of resistive control as a benchmark.



(a)

(b)

Figure 6.4: Convergence of the Q-value of each state action pair (a) and the absorbed energy per episode (b) with time in regular waves.



(a)

(b)

(c)

Figure 6.5: Time series of the (a) wave elevation ζ, body vertical displacement z and velocity $\dot{z}$, (b) PTO force F_{PTO} and generated power P in regular waves once the optimal policy has been found. Additionally, (c) shows the corresponding absorbed energy as compared with resistive control.

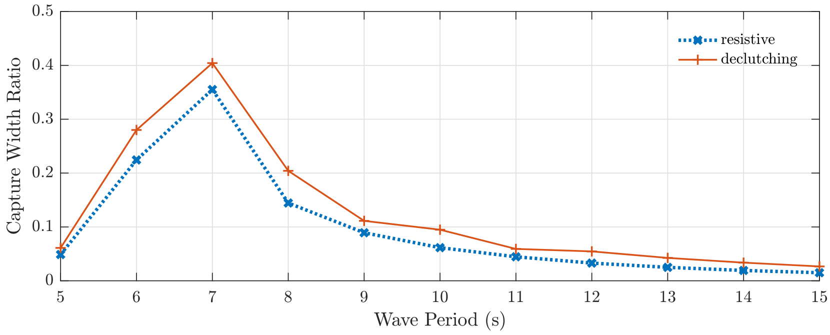


Figure 6.6: Variation in the capture width ratio of the point absorber with wave period in regular waves of unit amplitude for resistive and reinforcement learning declutching control.

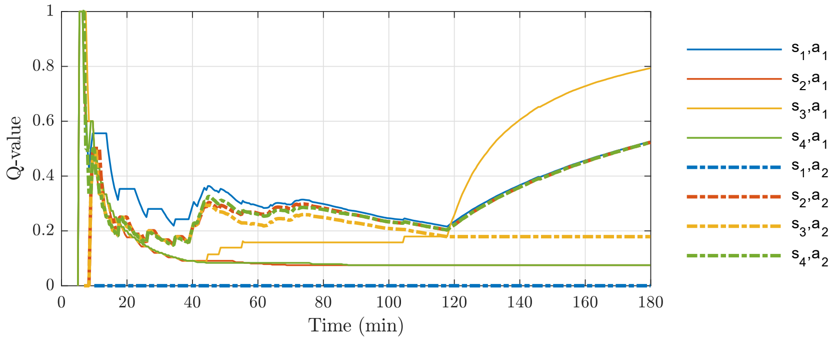


Figure 6.7: Convergence of the Q-value of each state action pair with time in irregular waves.

6.1.4 Results in irregular waves

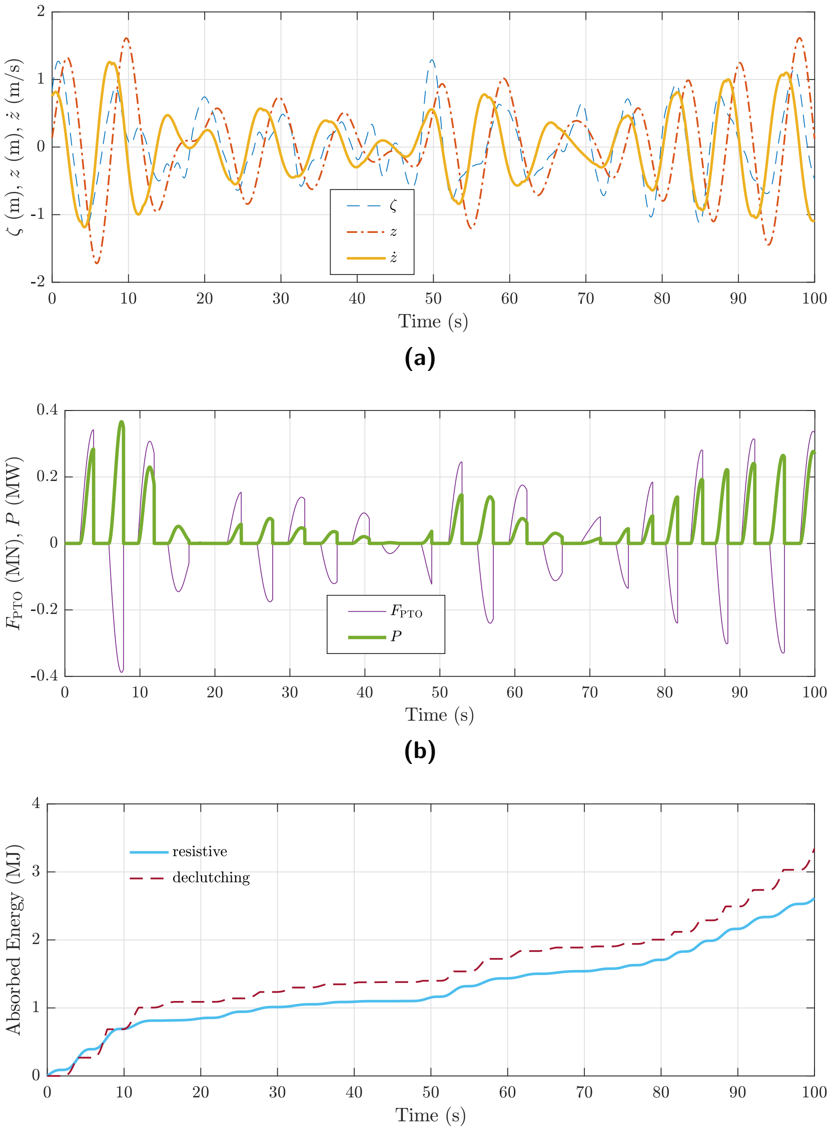
In Figure 6.7, it is possible to see the Monte-Carlo algorithm learning the optimal policy within three hours in the irregular wave trace.

The optimal policy is then shown in Fig. 6.8. The gain in absorbed energy associated with the Monte-Carlo declutching control over resistive control can be seen in Figure 6.8c.

6.1.5 Discussion of the results of Monte-Carlo declutching control

6.1.5.1 Regular waves

As it can be seen in Fig. 6.4b, the Monte-Carlo algorithm learns the optimal policy for the maximization of energy absorption in regular waves. It is possible to notice that the zig-zag nature of the curve over approximately the first 40 minutes is due to the predefined exploration of all possible policies. Figure 6.4a shows that the state-action pairs belonging to the optimal policy present the highest Q-values. Because of the nature of the return function as described in Section 6.1.1.3, it is interesting to notice that if the optimal policy is selected for a very large number of times once learning has



(a)

(b)

(c)

Figure 6.8: Time series of the (a) wave elevation ζ, body vertical displacement z and velocity $\dot{z}$, (b) PTO force F_{PTO} and generated power P in irregular waves once the optimal policy has been found. Additionally, (c) shows the corresponding absorbed energy as compared with resistive control.

finalised, the corresponding state-action values will converge to 1.

Figure 6.5 shows the optimal policy graphically, which is very similar to the one found with optimal command theory in Babarit *et al.* (2009). Nevertheless, in this work, when the bypass valve is closed, the control force is achieved with resistive control rather than through a step input. Additionally, from Figure 6.5, it is clear that there is an increase in energy absorption over resistive control, but without incurring any reactive power flow. This is the main advantage over reactive control (which in fact results in even greater power extraction), since strong reactive power flows, i.e. switches in power from positive to negative power values, can seriously damage the PTO system (Falcão, 2008; Babarit *et al.*, 2009). Furthermore, in Figure 6.5a, it is possible to see the non-linear behaviour of the body velocity due to declutching control. For this reason, a very small time step has been used to prevent numerical instabilities.

Figure 6.6 shows the increase in performance with declutching control over resistive control in regular waves in absolute terms. Although the increase in power absorption is not as great as it would be expected with reactive or latching control, it is of a comparable magnitude to that obtained by Babarit *et al.* (2009). Nevertheless, this is obtained with a much simpler method for WEC developers to implement, which furthermore does not depend on any models of the device dynamics. Even greater power absorption could be obtained if the current wave profile could be measured, and the wave elevation used as a state variable, as assumed in other works.

As described in Section 6.1.3, for a number of wave periods a different optimal policy has been found. This is because of the different phasing between the body vertical displacement and velocity in these sea states. This further supports the adoption of the wave elevation as a state variable.

6.1.5.2 Irregular waves

Figure 6.7 shows that the same optimal policy is valid for both regular and irregular waves with $H_\mathrm{s} = 2$ m and $T_\mathrm{e} = 8$ s. Nevertheless, as compared with Figure 6.4a convergence is slower. In particular, a suboptimal policy is followed until after approximately 2 hours, after which the optimal policy is found. This shows the capability of Monte-Carlo methods to learn the optimal policy for an infinitely large number of samples.

It is known that realistic sea states usually last between 0.5 and 6 hours (Holthuijsen, 2007). Hence, three hours may in fact be too long for convergence, with a likely further time increase in the learning process due to variable generator efficiency in practice. This time may be slightly reduced by shrinking the exploration rate more rapidly, decreasing the number of times all states are explored (4 at the moment), or both. Furthermore, three hours per sea state represent a very short time as compared with the expected operational life of the WEC, which ranges between 20 to 25 years.

In Figures 6.8a and 6.8b, it is possible to see the random nature of irregular waves, which is the main cause for the longer learning time required as compared with deterministic regular waves. However, Figure 6.8c shows that Monte-Carlo declutching control still results in a gain in power absorption over resistive control, even though not as pronounced as in regular waves (Figure 6.5c).

6.1.5.3 Comparison with optimal command theory

The Monte-Carlo algorithm results in a very similar control policy to the one obtained from optimal command theory by Babarit *et al.* (2009). Nevertheless, RL presents a much lower computational cost: the algorithm is updated only every $4T_\mathrm{e}$, with the update step itself requiring only minimal resources due to the small number of states and the nature of the equations used. Conversely, with optimal declutching control (Babarit *et al.*, 2009), at each time step, an optimization problem is solved in order to find the optimal control setting for the maximization of the energy absorption over a future time horizon, typically 3- to 10-s-long. Depending on the time step length (usually 0.005-0.1 s), each computation can be very resource-intensive. It is clear that if the optimization takes longer than the time step length, these algorithms cannot be implemented in practice.

However, the main advantage of the Monte-Carlo method proposed in this work over optimal command theory for the declutching control of WECs is the fact that it does not rely on knowledge of the instantaneous wave elevation. Yet, if this information were available, either through the use of wave buoys (e.g. Li *et al.* (2012)) or system identification (e.g. Giorgi *et al.* (2016b)), then a great increase in power absorption is to be expected with optimal command theory. Similarly, the forecast of the wave elevation can be incorporated within the reinforcement learning algorithm as an additional state variable.

Furthermore, real-time strategies such as optimal declutching control enable the inclusion of constraints not only on the PTO force, but also on the maximum allowable displacement in order to prevent damage when the end stops are reached. Conversely, with reinforcement learning strict displacement constraints cannot be enforced, but penalties can be used to teach the controller to avoid selecting a policy that results in excessive motions.

Instead of extending the initial work on declutching control, with the inclusion of the wave elevation as state variable and the addition of penalties on large displacements, we have preferred to apply reinforcement learning, and in particular the more effective temporal difference methods, to the resistive and reactive control of point absorbers.

This was done to reflect the state-of-the-art of the wave energy industry (based on direct experience at Pelamis Wave Power Ltd), with reinforcement learning expected to provide a more robust method for the application of existing control strategies. These developments are treated in the following sections.

6.1.5.4 Comments on the application of temporal difference methods to the declutching control of WECs

Based on the literature review in Chapter 4, initially Q-learning and Sarsa were selected for the application of reinforcement learning to the declutching control of the point absorber. In particular, the same state variables, action space and reward function as described in this section were adopted. However, it was soon discovered that these strategies were unable to converge towards the optimal policy. Although they did find the optimal policy in some instances, for other initial conditions this was not the case. Thus, the algorithms developed from Q-learning and Sarsa were not deemed robust and were discarded from practical applications. Monte-Carlo methods have proven to be much more robust. The reason for this is believed to be their underlying averaging nature.

Temporal difference methods may be applied if the state variables are modified so as to include the time to declutch.

6.2 Application of reinforcement learning to resistive control

In the application to resistive, or passive, control of the point absorber, reinforcement learning should be used to determine the optimal PTO damping coefficient in each sea state without relying on a model of the system dynamics. This approach is practical and efficient, and has been inspired by the work of Wei *et al.* (2015) and Wei *et al.* (2016) for the adaptive optimal control of wind turbines. If the PTO coefficient is discretized in a number of values, then the problem can be visualized as the extension of the grid-world in Section 4.6.1 to multiple dimensions, where the additional dimensions are provided by the sea state, i.e. the combination of statistical measures of the wave height and period.

The reinforcement learning formulation for the resistive control of the point absorber is summarized in Figure 6.9. At each step, the controller selects a change in B_{PTO}, the action, which is implemented by the PTO unit, the agent. How this is achieved in practice is dependent on the PTO type. The change in coefficient results in a reward that is a function of the generated power and in a change of state, where each state is represented by one value for the significant wave height, H_{s}, the mean zero-crossing or energy wave period, T_{z} or T_{e}, respectively, and the PTO damping coefficient.

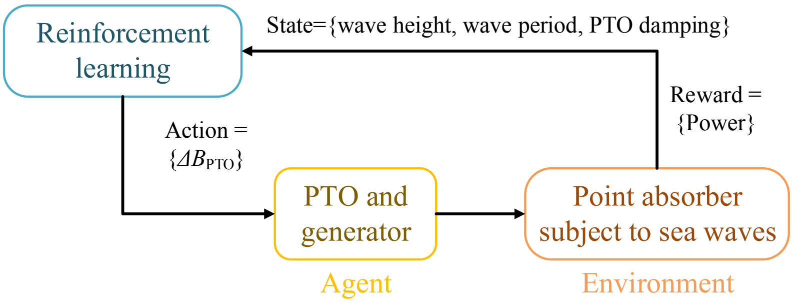

Reinforcement learning
State={wave height, wave period, PTO damping}
Reward = {Power}
Action = {ΔB_{PTO}}
PTO and generator
Point absorber subject to sea waves
Agent
Environment

Figure 6.9: Block diagram of the reinforcement learning control of the point absorber.

Due to the oscillatory nature of gravity waves, the generated power in the reward function needs to be averaged over at least one wave cycle. The averaging is performed over a horizon, H_{RL}, during which the state and action are constant, so that all time steps now have length H_{RL}. Then, a new action is selected, which results in an immediate change of state and a new averaging process.

The state-action value function is updated either at the end of each step, if an on-line reinforcement learning scheme is used, e.g. Q-learning and Sarsa, or off-line after a number of samples are collected, when LSPI is employed. These algorithms are described in Chapter 4.

The state and action spaces, reward function, learning and exploration rates, and discount factor of the reinforcement learning formulation of the resistive control are described in detail in the following sections.

6.2.0.1 State variables

As shown in Figure 6.9, the state variables are taken to be the significant wave height, either the energy wave period, and PTO damping coefficient so that the adopted reinforcement learning state space is:

$$\mathcal{S} = \left\{ s | s_{i,k,l} = \left(H_{\text{s},i}, T_{\text{e},k}, B_{\text{PTO},l} \right), \begin{array}{r} i = 1:I, \\ k = 1:K, \\ l = 1:L \end{array} \right\}. \tag{6.6}$$

If the mean zero-crossing or peak wave period are employed instead, T_{e} should be subsituted with T_{z} or T_{p}, respectively.

If discrete states (or tabular features) are used, the total number of features is given by $J = IKL$ and the state-action value function is exact. With function approximation, a smaller number of features may be adopted to reduce the computational cost. In fact, for the control of WECs, a hybrid approach is selected, where discrete sea states are still employed, while RBFs approximate the control variable. I and K are determined from the wave data at the deployment site, with steps of 0.5 m or 1 m, and 0.5 s or 1 s being common for the significant wave height and energy wave period, respectively (Holthuijsen, 2007). With a hydraulic PTO system, the value of L can be determined by the number of accumulators. Indeed, as shown in Henderson (2006), the time series of the PTO force is characterized by a number of discrete values.

6.2.0.2 Action space

Considering the selected state space, for passive control the action space is thus defined as:

$$\mathcal{A} = \{a|(-\Delta B_{\text{PTO}}, 0, +\Delta B_{\text{PTO}})\}, \tag{6.7}$$

where $\Delta B_{\text{PTO}} = B_{\text{PTO}l+1} - B_{\text{PTO}l}$. The states corresponding to the minimum or maximum damping coefficient, i.e. $B_{\text{PTO}1}$ and $B_{\text{PTO}L}$, have a limited (from 3 to 2) number of actions in order to prevent the controller from exceeding the state space boundary. For instance, for $B_{\text{PTO}1}$, the action $-\Delta B_{\text{PTO}}$ is precluded in the current state.

6.2.0.3 Reward function formulation

The reward function represents the goal that the controller is expected to maximise. Hence, for the passive control of WECs, the reward function needs to be a function of the absorbed power. Although Wei *et al.* (2015) and Wei *et al.* (2016) formulate the reward as a function of the *change* in power between time steps due to the legacy of the wind control industry (for instance as is practice with the maximum point tracking algorithm), the current power value should be used as reward instead. This is because reinforcement learning maximises the total reward, which is a function of both the current and expected future reward (through the state-action value), as described in Chapter 4. Hence, setting the change in power as reward function would result in the controller selecting those actions that result in the greatest change in power absorption. The region where the curve of absorbed power with PTO coefficient is steepest is unlikely to coincide with the region corresponding to the maximum extracted energy.

Even after recognizing that the reward should be a function of the mean generated power, P_{avg}, it is clear that this value may be more influenced by changes in the significant wave height than variations in the PTO damping coefficient depending on the chosen state refinement. This can be dealt with by using $P_{\text{avg}}/{H_{\text{s}}}^2$ as a reward, since the absorbed power is proportional to the square of the significant wave height (Holthuijsen, 2007). In addition, due to the discretization of the state variables and the stochastic nature of irregular waves, not only should the generated power be averaged over a horizon H_{RL} to produce P_{avg}, but the reward function needs to be built on the mean of a number M of these values for each state. This can be achieved by storing the M most recent $P_{\mathrm{avg}}/{H_{\mathrm{s}}}^2$ values for each discrete state (even when function approximation is employed) in a multi-dimensional list, $\boldsymbol{R}$. The size of the list is at most $(|\mathcal{S}|, M)$, with $|\mathcal{S}|$ being the number of discrete states. It is then possible to obtain the mean value in each discrete state and express it with the vector $\boldsymbol{m} = \langle \boldsymbol{R}(s_{\mathrm{d}}, m) \rangle_{m=1:(M \vee \mathrm{end})}$ of size $(|\mathcal{S}|, 1)$.

Depending on the selection of ΔB_{PTO}, the curve of the mean generated power with PTO coefficient can be very flat, particularly in the region near the optimal B_{PTO} value. This can cause the learning process to become very slow, since the benefit of picking the optimal damping coefficient in each sea state should be evident. This problem can addressed by raising the values within $\boldsymbol{m}$ to a power. This way the difference in mean generated power corresponding to neighbouring PTO coefficients is maximized. In order to prevent numerical instabilities, it is advantageous to first normalize the value of the vector for each state with the maximum value for each sea state. Hence, for discrete state s_{d}, the maximum value needs to be searched between the indices $o = \text{floor}\left(\frac{s_{\mathrm{d}}-1}{L}\right) L + 1$ and $p = \text{floor}\left(\frac{s_{\mathrm{d}}-1}{L}\right) L + L$ of the vector $\boldsymbol{m}$. The indices o and p ensure that the normalization is performed only over the values of $\boldsymbol{m}$ corresponding to the current sea state. The power value, defined as u, needs to be an odd number in order to avoid rectifying negative power values. Its determination is case specific.

In addition, in extreme seas the selected optimal damping coefficient may result in excessive motions, which can result in complete submergence or emergence of the machine or in the hit of the end stops (if present). This may cause severe structural damage if not complete failure. In order to prevent this, a penalty, $p < 0$, is returned when the magnitude of the maximum displacement over the averaging horizon exceeds a set value, z_{Max}. Usually, z_{Max} should represent a soft constraint in order to avoid failure. This is discussed in greater detail in Section 6.3, since greater displacements are a graver problem with reactive control.

The formulated reward function is thus expressed as follows for time step h (with duration H_{RL}):

$$r = \begin{cases} \left[\dfrac{\boldsymbol{m}\left(s_h\right)}{\max_{s''=o:p} \boldsymbol{m}\left(s''\right)}\right]^u & \text{if constraints met,} \quad (6.8a) \\ p & \text{otherwise.} \quad (6.8b) \end{cases}$$

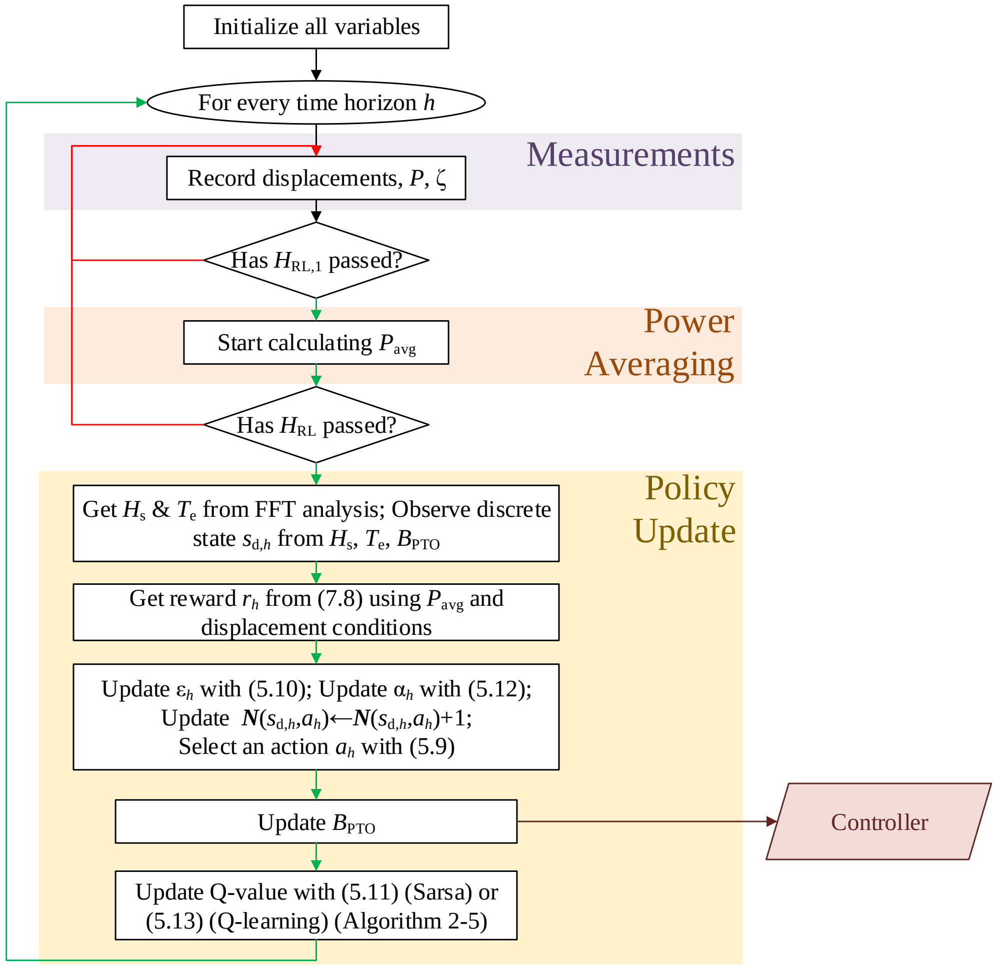

Initialize all variables
For every time horizon h
Measurements
Record displacements, P, ζ
Has $H_{\mathrm{RL},1}$ passed?
Power Averaging
Start calculating P_{avg}
Has H_{RL} passed?
Policy Update
Get H_s & T_e from FFT analysis; Observe discrete state $s_{\mathrm{d},h}$ from H_s, T_e, B_{PTO}
Get reward r_h from (7.8) using P_{avg} and displacement conditions
Update ε_h with (5.10); Update α_h with (5.12); Update $\boldsymbol{N}(s_{\mathrm{d},h},a_h)\leftarrow\boldsymbol{N}(s_{\mathrm{d},h},a_h)+1$; Select an action a_h with (5.9)
Update B_{PTO}
Controller
Update Q-value with (5.11) (Sarsa) or (5.13) (Q-learning) (Algorithm 2-5)

Figure 6.10: Flowchart of the Q-learning or Sarsa algorithm for the resistive control of the point absorber.

6.2.1 Algorithm

The algorithm of resistive control of WECs with reinforcement learning is summarised in the following two sections for on- and off-line strategies.

6.2.1.1 Q-learning and Sarsa

Figure 6.10 shows the algorithm for the resistive control of the point absorber using either Q-learning or Sarsa. If discrete states are used, the scheme should be built on top of Algorithms 3 and 2. When function approximation is employed, Algorithms 5 and 4 should be preferred instead.

As can be seen in Figure 6.10, the first stage of the algorithm is the initialization of all required variables. This includes the state-action value as well as the matrix $\boldsymbol{N}$, which contains the record of the number of visits to each discrete state-action pair. The value of L for the specification of the size of the matrix $\boldsymbol{R}$ has been set to 10 in regular waves and 25 in irregular waves. It is possible to speed up convergence by pre-calculating the entries of the $\boldsymbol{R}$ matrix in a run in a similar wave trace, whilst taking random actions. Simulations can also be used to initialize the $\boldsymbol{R}$ matrix for the full-scale device, since its entries will slowly be replaced from those of the actual environment.

After the initialization phase, the algorithm is run indefinitely until maintenance is due. At every time step, the generated power, the displacements and the wave elevation are measured. The mean generated power is computed only over the latter part of each averaging time period, i.e. after a time of $H_{\mathrm{RL},1}$ has passed since the start of the horizon. This improves the learning behaviour by filtering the transient effects associated with a change in PTO damping coefficient. At the end of the time interval, with duration H_{RL}, the actual policy update takes place according to Algorithms 2-5.

The selection of suitable values for H_{RL} is based on a compromise between a small value for quicker response and a large value for a more stable algorithm. Indeed, although a sea state can be stationary for a period ranging from 20 minutes to 6 hours (Holthuijsen, 2007), individual neighbouring waves within this time can present very different characteristics, with energy packets typically being carried by wave groups with similar energy content. Continuous changes in the sea state from a step to the next prevent the algorithm from converging, since the agent could land in any states due to the environmental noise. Therefore, although the aim should be to obtain a real-time control in the future, as discussed in Chapter 7, in this thesis a longer averaging time spanning multiple wave periods has been selected to ensure converge of reinforcement learning as a proof of concept. 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.

6.2.1.2 Least-squares policy iteration

Figure 6.11 shows the algorithm for the resistive control of the point absorber using LSPI. The algorithm is very similar to Figure 6.10. The main difference lies in the additional sample collection stage, with the policy improvement taking place off-line as described in Chapter 4. The policy is improved using the LSPI algorithm in Algorithm 7 every $N_h = 40$ time horizons. This operation can be performed on separate computing cores so as to reduce the computational effort and ensure the on-line implementation is feasible.

At the end of each horizon, the current state, action, next state and reward are sampled and added to $\boldsymbol{S}$. 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

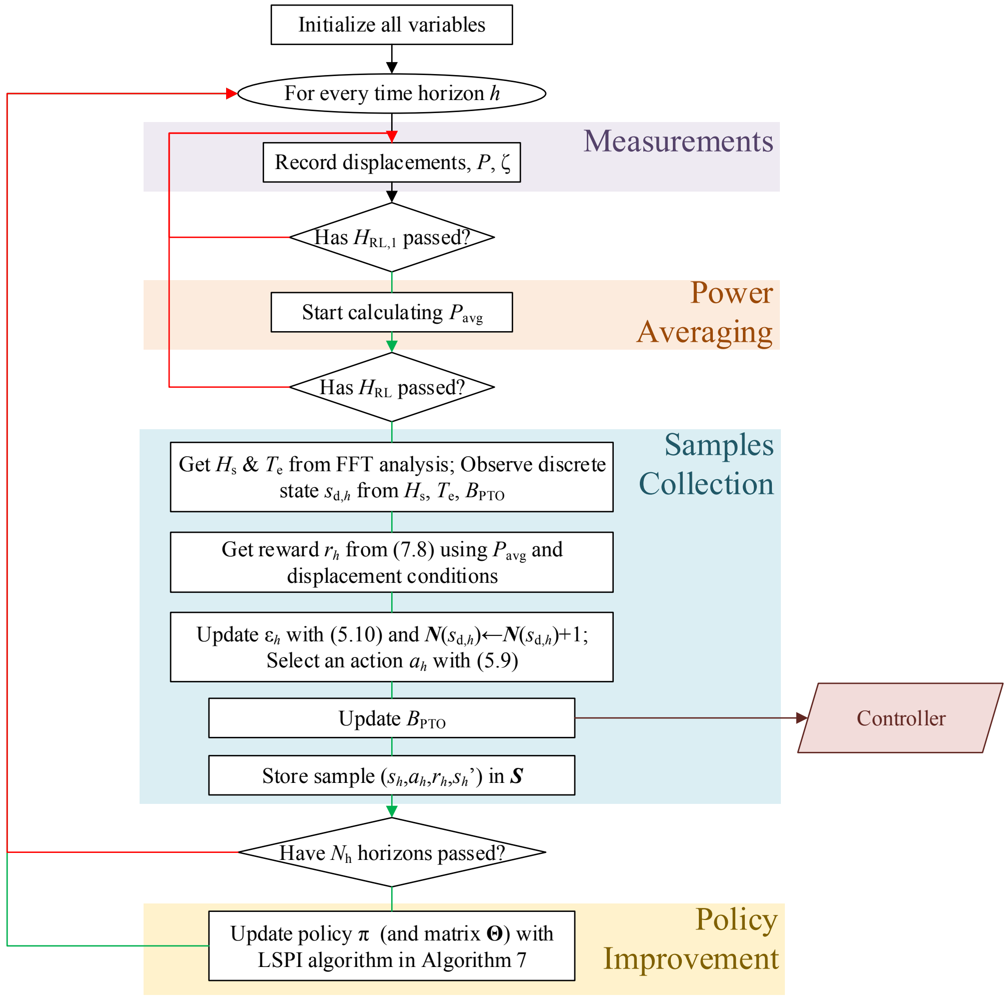

Initialize all variables
For every time horizon h
Measurements
Record displacements, P, ζ
Has $H_{\mathrm{RL},1}$ passed?
Power Averaging
Start calculating P_{avg}
Has H_{RL} passed?
Samples Collection
Get H_{s} & T_{e} from FFT analysis; Observe discrete state $s_{\mathrm{d},h}$ from H_{s}, T_{e}, B_{PTO}
Get reward r_h from (7.8) using P_{avg} and displacement conditions
Update ε_h with (5.10) and $\boldsymbol{N}(s_{\mathrm{d},h})\leftarrow\boldsymbol{N}(s_{\mathrm{d},h})+1$; Select an action a_h with (5.9)
Update B_{PTO}
Controller
Store sample (s_h,a_h,r_h,s_h') in $\boldsymbol{S}$
Have N_{h} horizons passed?
Policy Improvement
Update policy π (and matrix $\boldsymbol{\Theta}$) with LSPI algorithm in Algorithm 7

Figure 6.11: Flowchart of the LSPI algorithm for the resistive control of the point absorber.

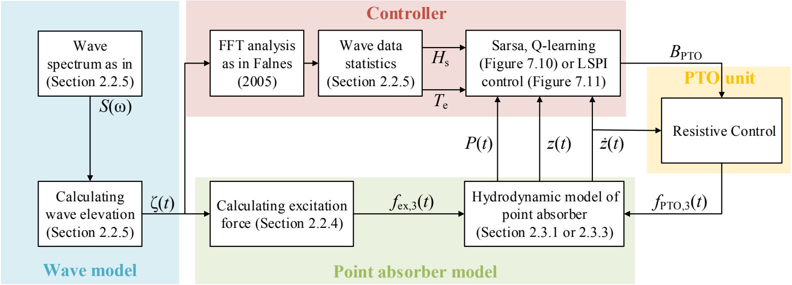

Controller
Wave spectrum as in (Section 2.2.5)
$S(\omega)$
FFT analysis as in Falnes (2005)
Wave data statistics (Section 2.2.5)
H_s
T_e
Sarsa, Q-learning (Figure 7.10) or LSPI control (Figure 7.11)
B_{PTO}
PTO unit
Resistive Control
$P(t)$
$z(t)$
$\dot{z}(t)$
Calculating wave elevation (Section 2.2.5)
$\zeta(t)$
Calculating excitation force (Section 2.2.4)
$f_{ex,3}(t)$
Hydrodynamic model of point absorber (Section 2.3.1 or 2.3.3)
$f_{PTO,3}(t)$
Wave model
Point absorber model

Figure 6.12: Work-flow diagram of the program used to simulate the point absorber with reinforcement-learning-based resistive control.

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

6.2.2 Case studies

The proposed resistive control for WECs based on reinforcement learning has been tested on two numerical benchmark cases:

- the linear point absorber with a single degree of freedom, first introduced in Section 2.3.1;
- the model of the Seabased device, which includes non-linearities in the PTO model, as described in Sections 2.3.3 and 3.2.1.2.

The first case study represents a benign framework for a proof of concept. The example is used to show the learning behaviour of the control even in multiple sea states in irregular waves. In addition, the behaviour of the controller in case of PTO force saturation is described. The second case study includes the influence of non-linear effects and end stops in order to study their impact on the convergence properties of the algorithm. Furthermore, the adaptiveness of the control is assessed by changing the system dynamics as if due to marine bio-fouling.

In both case studies, the simulation system shown in Figure 6.12 has been used.

6.2.3 Case study: linear point absorber

First of all, the resistive control based on reinforcement learning is applied to the point absorber restricted to heaving motions introduced in Section 2.3.1. The maximum PTO force has been assumed to be 1 MN, while the float displacement has been limited to ±5 m (for the application of any penalties).

The PTO system of the device has been assumed to be composed of 4 accumulators. Hence, 9 discrete values are used for the PTO damping coefficient, which ranges from 0 to 800 kNs/m with $\Delta B_{\text{PTO}} = 100$ kNs/m. In this initial study, a smaller number of discrete states are used in irregular waves to speed up convergence when the control is tested in multiple sea states in random seas. In this case, only 5 damping coefficients values are employed, with the same range but $\Delta B_{\text{PTO}} = 200$ kNs/m. However, a wider range and finer resolution are likely to be required for a more realistic implementation.

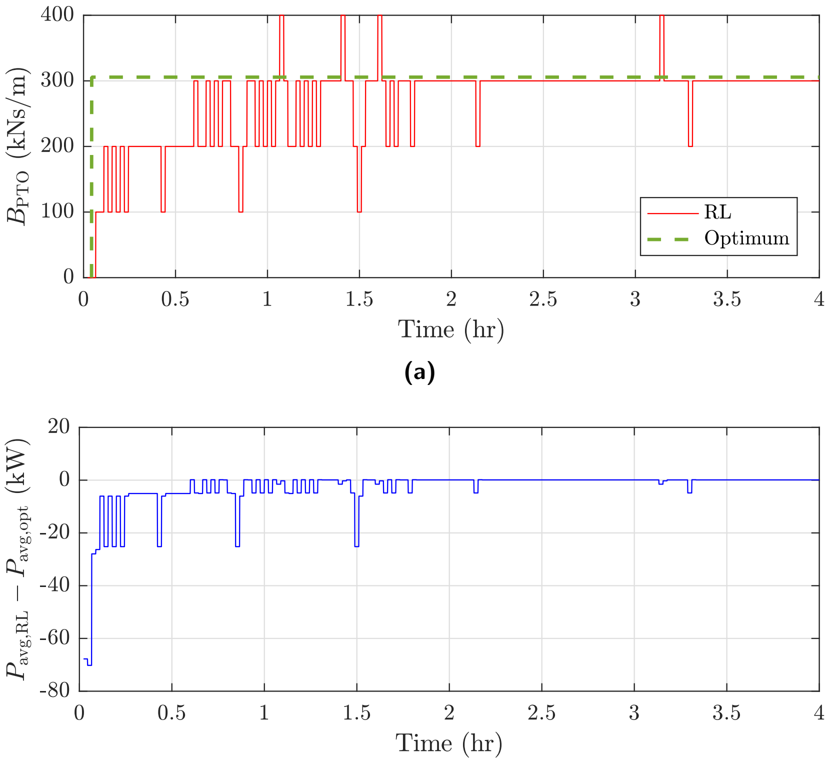
For this case study, only Q-learning is investigated. The values of the discount factor, the initial exploration and learning rates are set to $\gamma = 0.75$, $\epsilon_0 = 0.5$ and $\alpha_0 = 0.4$, respectively. Furthermore, the parameters for the decay of the exploration and learning rates are specified as $N_\epsilon = 25$ and $N_\alpha = 5$, respectively. In the reward function, the values $u = 21$ and $p = -2$ have been adopted. The duration of each time horizon has been set to $H_{\text{RL}} = 10T$ in regular waves and $H_{\text{RL}} = 30T_{\text{z}}$ in irregular waves. These values have been found to be the minimum in order to ensure convergence. The time averaging process is started after $H_{\text{RL},1} = 5T_{\text{z}}$, which represents a heavy filter.

6.2.3.1 Results in regular waves

Regular waves have been analysed first in order to assess the convergence properties of the proposed reinforcement-learning-based control (RL) under deterministic conditions. A single sea state (hence, $I = K = 1$) with unit wave amplitude and a wave period of 8 s has been considered, with the time series lasting 4 hours.

Figure 6.13a shows the convergence of the reinforcement learning algorithm towards the optimal PTO damping for this sea state. The optimal value has been calculated with state-of-the-art resistive control as described in Section 3.2 using a wave trace lasting 20 minutes in the simulation. The corresponding mean absorbed power over $H = 10T_{\text{z}}$ can be seen in Figure 6.13b.

Due to the low wave height selected in all simulations, the PTO force never reaches its limit, with the maximum force being 237.910 kN for the optimal B_{PTO} in Figure 6.13. In order to analyse the effects of the force clip, or saturation, on the optimal PTO damping coefficient and the learning process, the force limit has been reduced to $F_{\text{Max}} =$ 237.910 kN. Then, the wave amplitude has been slightly increased to 1.1 m. This is analogous to the device reaching the original saturation limit in more extreme waves,



(a)

(b)

Figure 6.13: Optimal and reinforcement-learning-selected PTO damping coefficient (a) and corresponding mean absorbed power (b) in regular waves of unit amplitude and a wave period of 8 s.

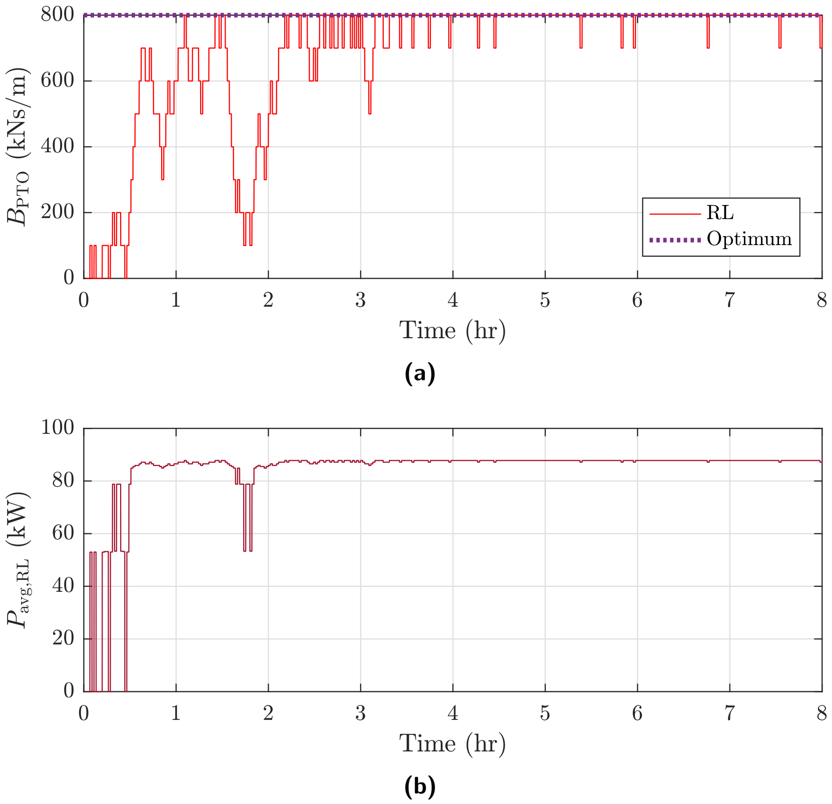


Figure 6.14: PTO damping coefficient selected by the Q-learning algorithm (a) and corresponding mean generated power (b) in regular waves with $H_s = 2.2$ m and $T_z = 8$ s, when $F_{\text{Max}} = 237.910$ kN.

whilst the validity of the assumption of linear wave theory in the hydrodynamic model is ensured.

The convergence of the reinforcement learning algorithm towards a new PTO damping coefficient and the corresponding mean absorbed power can be seen in Figures 6.14a and 6.14b respectively. Note that the optimal B_{PTO} value would be far beyond the state space we have defined, so that it is saturated at 800 kNs/m. The reason for this behaviour can be understood by looking at Figure 6.15, which shows the variation of the PTO velocity and force over time with the two different PTO damping coefficients, 300 and 800 kNs/m, in regular waves of unit amplitude and a wave period of 8 s. With the lower saturation limit $F_{\text{Max}} = 237.910$ kN, the controller tries to maximise the absorbed power by maximising the area under the curve of the PTO force through a square wave. The limit on the PTO damping coefficient prevents the realization of a fully non-linear, bang-bang type of control response.

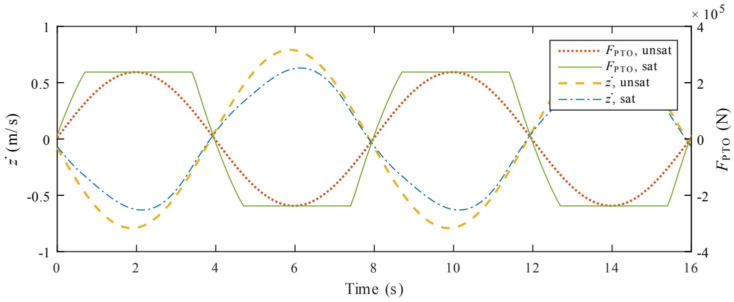


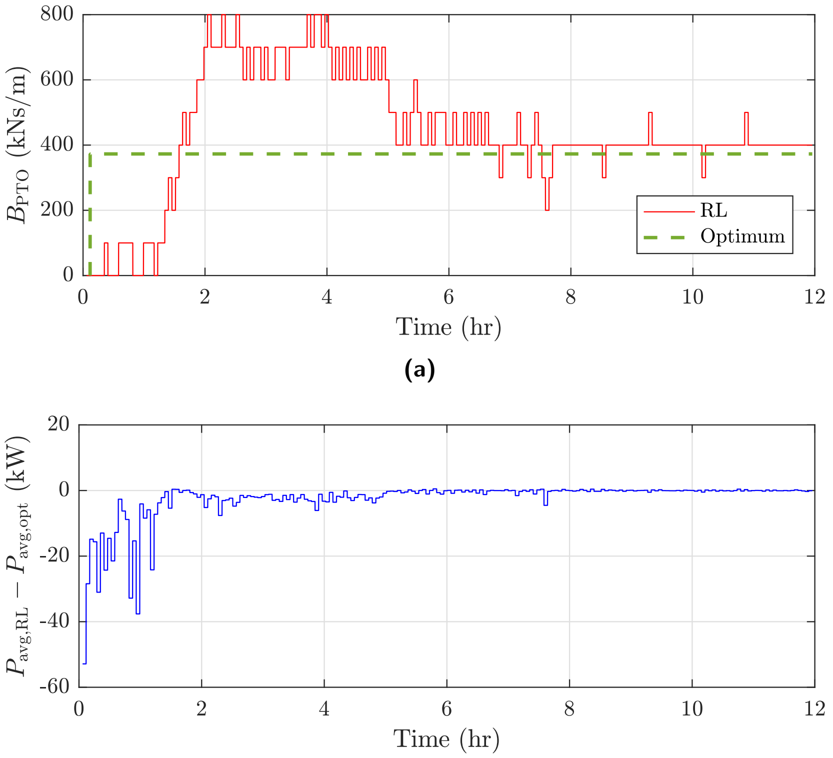
Figure 6.15: PTO velocity and force over two wave periods in regular waves with $H_{\text{s}} = 2$ m and $T_{\text{z}} = 8$ s for the cases of unsaturated ($B_{\text{PTO}} = 300$ kNs/m) and saturated ($B_{\text{PTO}} = 800$ kNs/m) PTO force, when $F_{\text{Max}} = 237.910$ kN.

6.2.3.2 Results in irregular waves

6.2.3.2.1 Single sea state

Firstly, a wave trace generated using a single JONSWAP spectrum is considered, with a significant wave height of 2 m and a peak wave period of 9 s, corresponding to $T_{\text{z}} = 7$ s from the spectral analysis. Although there are oscillations in the predicted values of H_{s} and T_{z} over neighbouring horizons, $J = K = 1$ have been used for simplicity, so that $J = 9$. The wave trace lasts 12 hours, with a preliminary initialization lasting 15 minutes.

Figure 6.16a shows the convergence of the PTO damping coefficient selected by Q-learning towards the optimum. The optimal value has been calculated as described in Section 3.2, where the simulations rely on a 20-minute-long wave trace with 5 different seed values. The mean absorbed power corresponding to the selected and optimal B_{PTO} values can be seen in Figure 6.16b.



(a)

(b)

Figure 6.16: Optimal and reinforcement-learning-selected PTO damping coefficient (a) and corresponding mean absorbed power (b) in irregular waves with $H_{\mathrm{s}} = 2$ m and $T_z = 7$ s, generated using a JONSWAP spectrum.

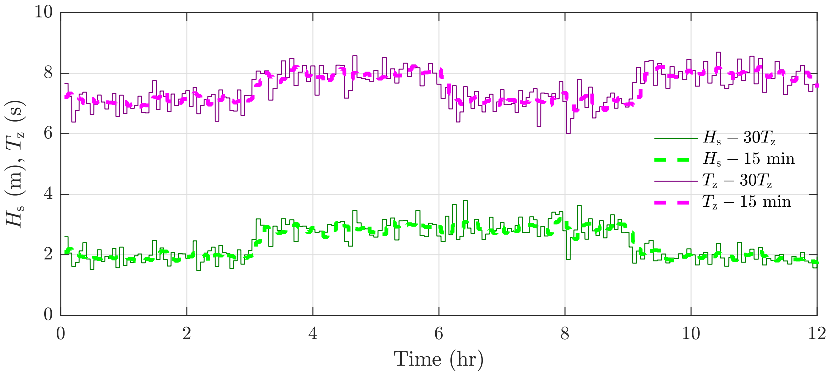


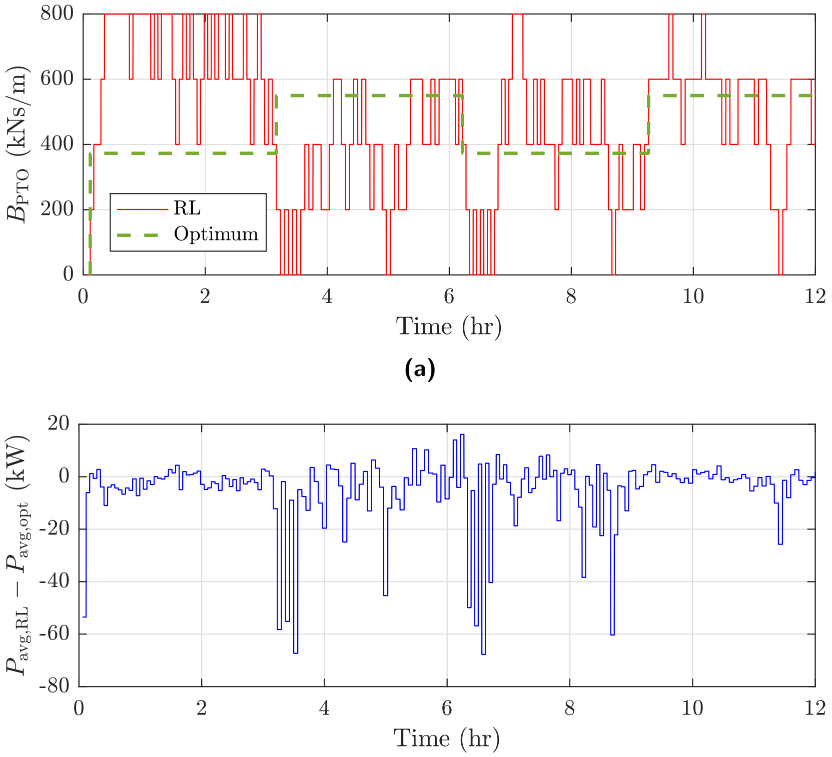
Figure 6.17: Significant wave height and mean zero-crossing period calculated over each horizon (continuous lines) and over overlapping 15-minute windows every minute (dotted lines) for the multiple sea state wave trace.

6.2.3.2.2 Multiple sea states

In ocean waves, sea states can last from a minimum of 30 minutes to a maximum of six to eight hours, with swells lasting typically between three and six hours (Holthuijsen, 2007). Hence, a semi-realistic wave trace (Figure 6.17) has been generated from the concatenation of four sea states, each lasting three hours and corresponding to a JONSWAP spectrum. In order to achieve convergence, the wave trace has been repeated 4 times for a total of 48 hours.

In Figures 6.19a and 6.19b, it is possible to see the behaviour of the Q-learning algorithm when $I = 1$ ($H_s = 2$ m) and $K = 4$ ($T_z = 6, 7, 8, 9$ s). Although only four wave spectra are employed to generate the sea state, determining the sea state over the horizon length H_{RL} results in 4 discrete values of both the significant wave height (1-4 m, in steps of 1 m) and the mean zero-crossing wave period (6-9 s, in steps of 1 s). However, of these only one value is employed for the significant wave height, 2 m, since the wave energy is too low for the generator to reach its limit within this wave trace. This means that the optimal damping coefficient is dependent only on the T_z value.

Figure 6.18 shows the initial behaviour of the Q-learning algorithm, while Figure 6.19 shows the control performance after the optimal PTO damping coefficient has been learnt in each sea state. Figures 6.18a and 6.19a also present the optimal value for the PTO damping coefficient, calculated as described in the previous section for the four individual sea states. However, as opposed to the reinforcement learning method, in this case the values of H_s and T_z are obtained from 15-minute moving windows every minute, as shown by the dotted lines in Figure 6.17. In Figures 6.18b and 6.19b, it is possible to see the difference in the mean absorbed power obtained using reinforcement learning and the optimal PTO damping coefficient, where the optimal mean absorbed power has an average value of 26.147 kW, 56.429 kW, 59.191 kW, and 25.208 kW in



(a)

(b)

Figure 6.18: Optimal and control-selected ($I = 1$, $K = 4$, $L = 5$) PTO damping coefficient (a) and corresponding mean absorbed power (b) in irregular waves with four sea states, generated from the combination of $H_{\mathrm{s}} = 2, 3$ m and $T_{\mathrm{z}} = 7, 8$ s.

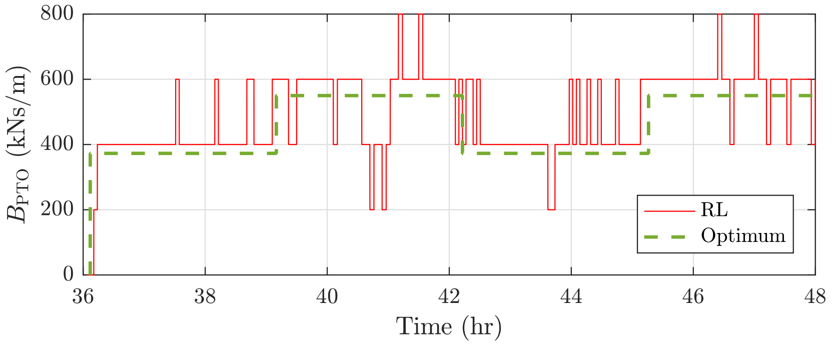
each sea state respectively, over the 3-hour wave traces.

6.2.3.3 Discussion

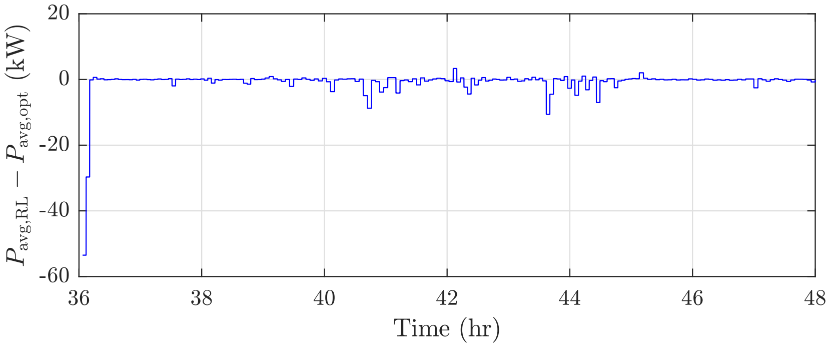
6.2.3.3.1 Regular waves

As can be seen from Figure 6.13, in regular waves the Q-learning algorithm can converge towards the optimal PTO damping coefficient for passive control in less than three hours starting from a random initialization. This is possible because of the deterministic nature of regular waves, which also enables the use of a relatively short averaging horizon. Similarly, the use of the tabular approach for the reward function would not be necessary. It is also interesting to notice that due to the selected exploration strategy, random actions may be taken even after the state-action values have fully converged.

From Figure 6.14, it is clear that the application of the force clip results in the optimal PTO damping coefficient moving to the upper limit. As aforementioned, the reason for this behaviour is the fact that the control force tends to a square wave shape (Figure 6.15), which maximises the area under the curve. Conversely, due to the force



(a)



(b)

Figure 6.19: Optimal and control-selected $(I = 1,\ K = 4,\ L = 5)$ PTO damping coefficient (a) and corresponding mean absorbed power (b) in irregular waves with four sea states, generated from the combination of $H_{\mathrm{s}} = 2, 3$ m and $T_{\mathrm{z}} = 7, 8$ s (Continuation of Figure 6.18).

saturation, the magnitude of the body velocity, which corresponds to the velocity at the PTO in this simple case, is not significantly affected by the PTO force. Since the absorbed power is proportional to the product of the PTO velocity and force, a square wave shape of the PTO force maximises the amount of generated energy. Hence, the controller is able to turn to a bang-bang type of control action when the force saturates, which can result in greater energy absorption than resistive control, as for instance shown by Li *et al.* (2012), despite a stronger generator loading. Nevertheless, as this section focuses on the application of reinforcement learning to resistive control, a relatively low limit has been imposed on the PTO damping coefficient to prevent the controller behaviour from becoming strongly non-linear.

In Figure 6.15, it is also interesting to notice that the saturated body velocity, like the PTO force, is no longer sinusoidal. The two curves are still in phase, but the velocity is affected by the higher order harmonics of the PTO force due to the saturation.

Similarly, although the specified limit on the body displacement is never reached in the tests considered, the reinforcement learning control would be expected to return a higher PTO damping coefficient than the optimal value if this were the case. Indeed, stronger damping is associated with a smaller motion amplitude.

6.2.3.3.2 Irregular waves

The statistical reward function is proven to be very effective in the treatment of irregular waves, as it is clear from Figure 6.16. However, a longer time is required for convergence to occur as compared with regular waves. This is evident from the comparison of Figures 6.18 and 6.19, which respectively show a random response while the controller is learning and the optimal performance once convergence is achieved. From this analysis of multiple sea states, it is possible to deduce that the controller needs to spend a minimum of 12 hours in each sea state in order to learn the optimal policy by ensuring sufficient exploration, when 5 values are used for the PTO damping coefficient (for a total of 20 states, not all of which are encountered). This time is likely to rise when a finer mesh is used for the reinforcement learning state space. In particular, assuming the learning time to be proportional to the number of states, a very large number of discrete B_{PTO} values can seriously affect the convergence properties of the algorithm, since the number of states is equal to the product of L and the number of sea states.

Although a 12-hour-long learning time seems much longer than the 20-minute window used for the Nelder-Mead optimization, multiple iterations are required for convergence with any search technique, so that reinforcement learning does in fact converge faster in an on-line application. In fact, a real-time, model-free implementation of an exhaustive search method would be impossible. Since in the real environment a wave trace is never repeated exactly, any search scheme would be unable to recognize whether a change in the cost function is due to the change in PTO damping or wave noise. Conversely, as Figure 6.19 shows, the proposed Q-learning strategy is able to start the optimization in any sea state from where it left off the last time it entered that specific sea state. Once convergence is achieved, the reinforcement learning approach is reduced to a look-up-table method until the exploration rate is increased in order to check if there have been any changes in the dynamics of the device. This can be done every season, but it will result in much shorter learning times during which the performance will never be far from the optimum, since the state-action value function is already initialized. Thus, as the operational life of a WEC is planned as 25 years, a relatively poor efficiency during the very first stages of operation should not affect the economic performance of the device.

From Figure 6.19a, it may look like the Q-learning algorithm has still not fully learnt the optimal policy even after 48 hours, despite a much better performance as compared with Figure 6.18a. In fact, the state-action value function has by now converged towards the correct optimal PTO damping coefficient in each sea state. However, the optimal values in the $\boldsymbol{m}$ vector, used to calculate the reward function, lie closest to $B_{\text{PTO}} = 200$ kNs/m for $T_{\text{z}} = 6$ s, $B_{\text{PTO}} = 400$ kNs/m for $T_{\text{z}} = 7$ s, $B_{\text{PTO}} = 600$ kNs/m for $T_{\text{z}} = 8$ s, and $B_{\text{PTO}} = 800$ kNs/m for $T_{\text{z}} = 9$ s. Hence, the oscillations in the PTO damping coefficient selected by the Q-learning algorithm in fact correspond to changes in sea state, as it is possible to understand from a close comparison with Figure 6.17. As a result, the reinforcement learning method even presents higher power absorption at some points as compared with the standard resistive control in Figure 6.19b despite the use of a very coarse reinforcement learning state space at this stage.

No comparison has been made at this stage with other control strategies, such as latching or model predictive control, because reinforcement learning is considered to be a method to make existing control strategies independent of the hydrodynamic model of the wave energy converter. Hence, its performance is only as good as the control scheme itself.

6.2.3.4 Summary

This simple case study has shown that the proposed Q-learning algorithm is able to learn the optimal PTO damping coefficient in each sea state for the resistive control of a WEC. Furthermore, the controller can recognize the current sea state and pick up learning from where it left off the last time the sea state was encountered. Nevertheless, despite the linearity of the model, up to 12 hours are required for convergence in each sea state. In spite of the long expected life time of a WEC, this value is considered to be still excessive. For this reason, in the following case study, we investigate whether

Table 6.2: 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

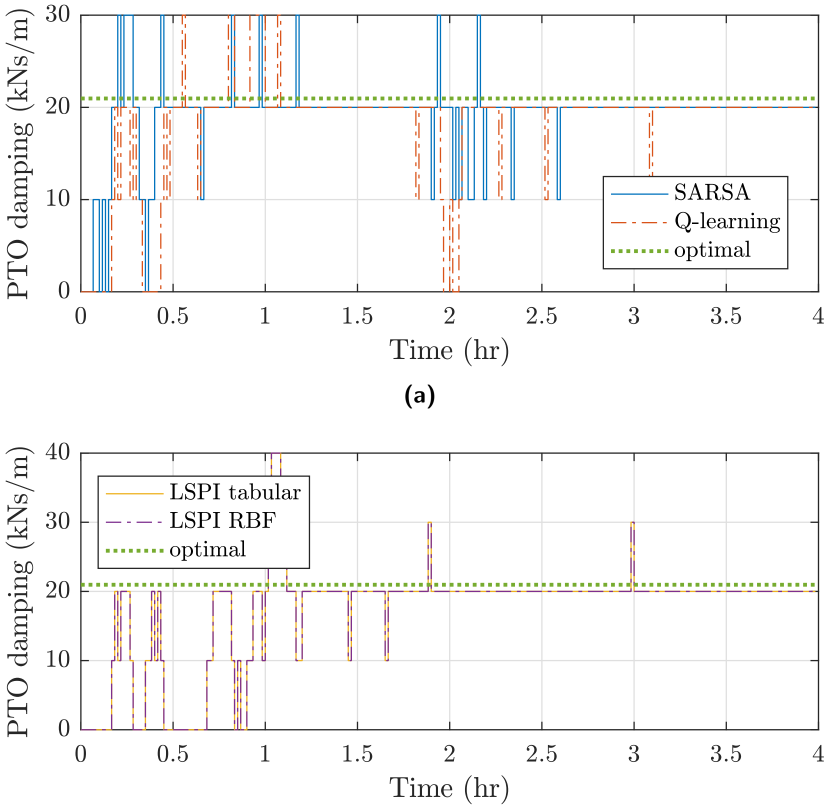
function approximation and more efficient algorithms, such as LSPI, can help reduce the learning time. Additionally, the claim of controller adaptability is studied in detail.

6.2.4 Case study: Seabased device

The second case study considers the model of the Seabased WEC described in Sections 2.3.3 and 3.2.1.2 ((2.96)-(2.103)), which has non-linear mooring (due to the slacking of the mooring line) and PTO (due to the end stops) forces. This device presents end-stops, which means that in the reward function a penalty $p = -1$ is returned if $\max y > l_\mathrm{u} + l_\mathrm{e,u}$ or $\min y < -(l_\mathrm{l} + l_\mathrm{e,l})$ during each time horizon, where y is the displacement of the translator. The PTO damping coefficient has been assumed to range from 0 to 100 kNs/m in steps of 10 kNs/m (for a total of 11 discrete values per sea state) based on preliminary calculations.

For this case study, three reinforcement learning algorithms are assessed: Q-learning, Sarsa and LSPI. In addition, for LSPI both tabular and radial features have been analysed. With RBFs, a smaller number of values can be used. In particular, five 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 6.2, 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 (4.17). The first kernel is always sited at $\gamma = 0$ kNs/m.

The values of the discount factor and initial exploration and learning (for Q-learning and Sarsa only) rates have been set to $\gamma = 0.95$, $\epsilon_0 = 0.5$ and $\alpha_0 = 0.4$, respectively. Furthermore, the parameters for the decay of the exploration and learning (for Q-learning and Sarsa only) rates are specified as $N_\epsilon = 5$ and $N_\alpha = 5$, respectively. In the reward function, a power value of $u = 25$ has been adopted. The duration of each time horizon has been set to $H_\mathrm{RL} = 10T$ in regular waves and $H_\mathrm{RL} = 25T_\mathrm{e}$ in irregular waves. These values have been found to represent the best compromise between convergence speed and computational cost. The time averaging process is started after $H_\mathrm{RL,1} = 5T_\mathrm{e}$, which represents a heavy filter. With LSPI, the policy is improved every $N_h = 40$ samples.



(a)

(b)

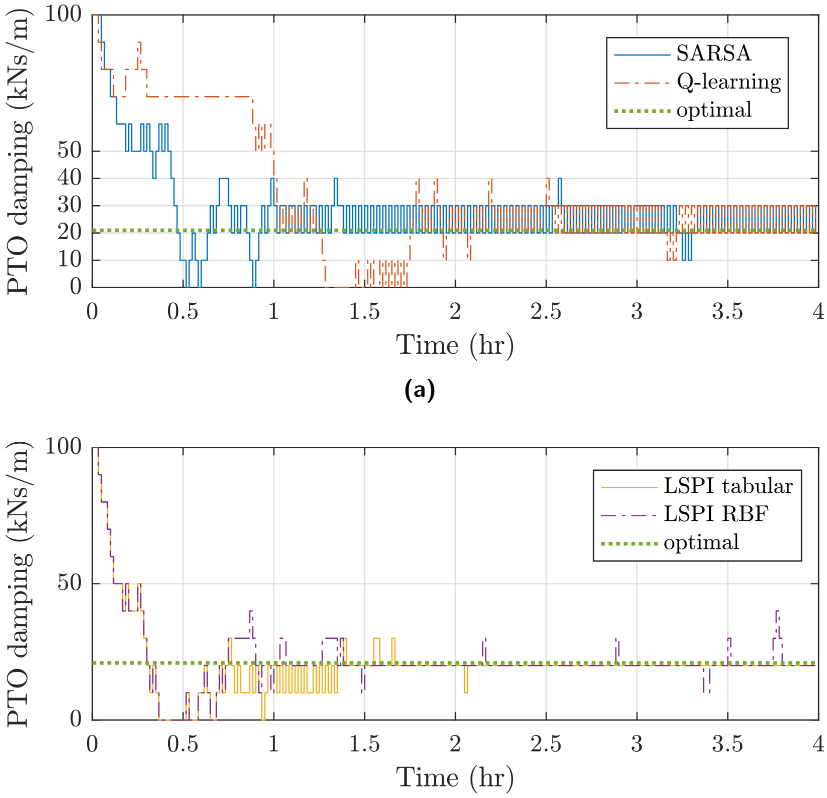
Figure 6.20: PTO damping coefficient selected by different reinforcement learning control strategies as compared with the optimal value in regular waves with unit amplitude and $T = 6$ s starting from $B_{\mathrm{PTO}} = 0$ kNs/m.

6.2.4.1 Results in 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 as described in Section 3.2. The optimizer is run with a time-domain simulation in a 20-minute-long wave trace 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 $B_{\mathrm{PTO}} = 0$ and $B_{\mathrm{PTO}} = 100$ kNs/m, as shown in Figures 6.20 and 6.21, 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 Figure 6.21 in order to assess the convergence properties under difficult conditions.

In Figure 6.22, it is possible to see the behaviour of the LSPI algorithm for the RBF



(a)

(b)

Figure 6.21: PTO damping coefficient selected by different reinforcement learning control strategies as compared with the optimal value in regular waves with unit amplitude and $T = 6$ s starting from $B_{\text{PTO}} = 100$ kNs/m.

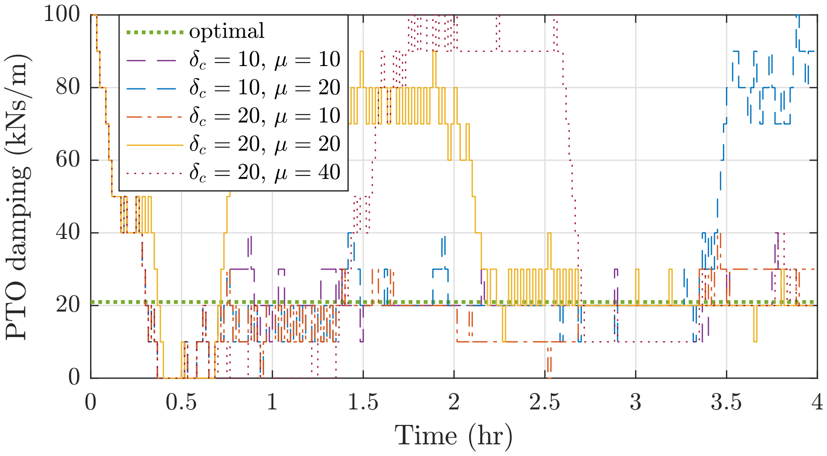


Figure 6.22: PTO damping coefficient selected by the LSPI algorithm with different RBF settings in regular waves with unit amplitude and $T = 6$ s. The values of δ_c and μ are in kNs/m.

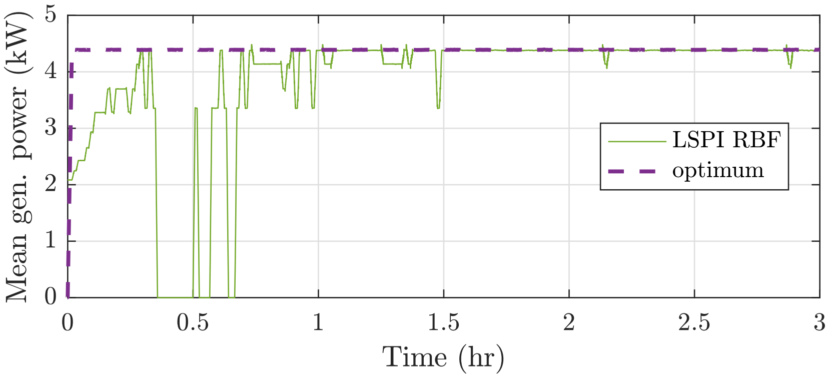


Figure 6.23: Mean generated power for the run with LSPI with RBFs and $\delta_c = 10$ kNs/m and $\mu = 10$ kNs/m in Figures 6.21b and 6.22.

settings in Table 6.2, when the starting value of the PTO coefficient is $B_{\text{PTO}} = 100$ kNs/m. For all runs, the same seed values is used as in Figure 6.21. 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 Figures 6.21b and 6.22 is plotted in Figure 6.23.

6.2.4.2 Results in irregular waves

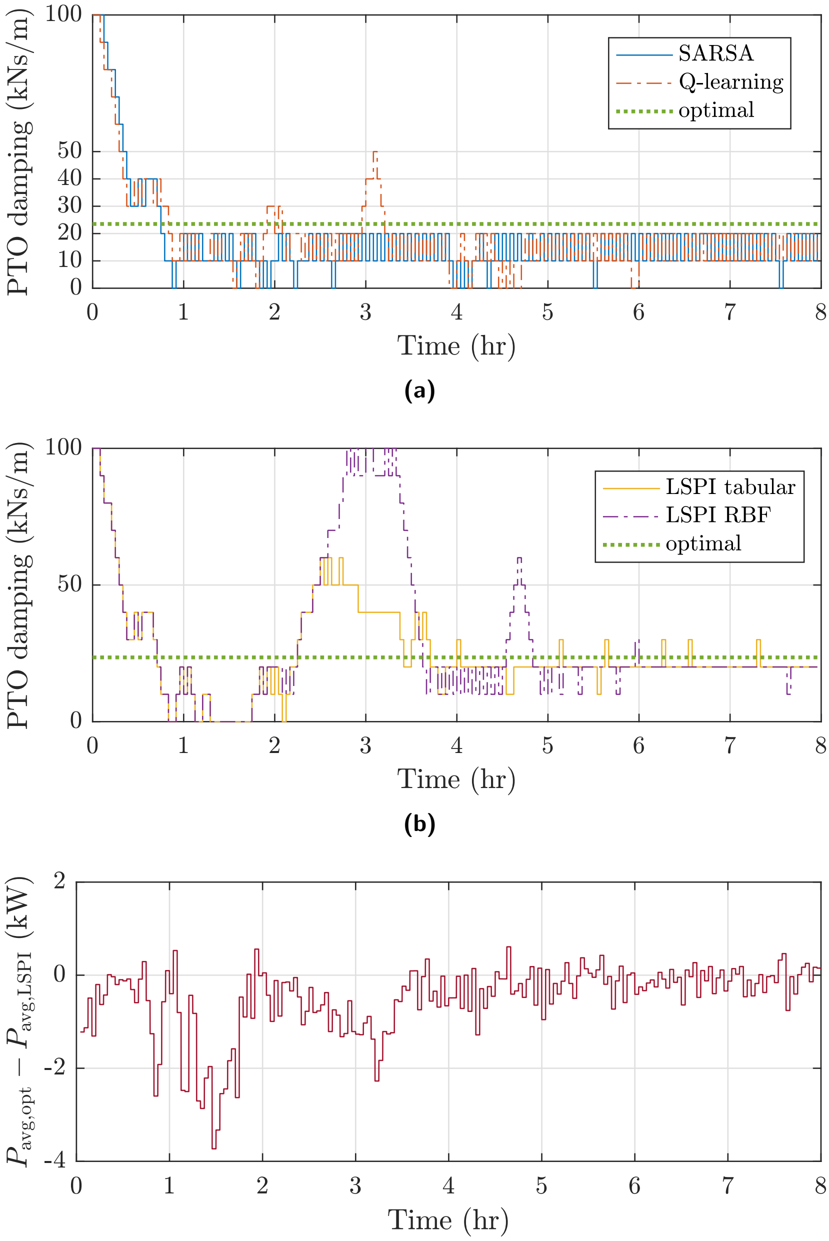
In irregular waves, an 8-hour long wave trace with $H_\mathrm{s} = 2$ m and $T_\mathrm{e} = 6$ s with a JONSWAP spectrum has been analysed, typical of the Lysekil testing site (Waters *et al.*, 2009). In Figures 6.24a and 6.24b, the learning behaviours of the three control algorithms are shown, with the same setting being used for LSPI with RBFs as in Figure 6.21 throughout this section. The difference in mean generated power between LSPI with RBFs and the optimal control setting is shown in Figure 6.24c.

Nevertheless, real sea states actually last between 0.5 to 6 hours (Holthuijsen, 2007). Therefore, in order to prove that reinforcement learning 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 states, so that $I = K = 2$. 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. Figures 6.25a and6.25b show the learning behaviour of the three reinforcement learning algorithms. In Figure 6.25c, the difference in mean power between LSPI with RBFs and the optimal control setting in each sea state can be seen.

Furthermore, although reinforcement learning is expected to result in adaptive control, as it is model-independent (Lewis *et al.*, 2012), this was not proven in the previous case study. Hence, a simple example is treated here to show the adaptivity of reinforcement learning 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 Eriksson *et al.* (2007)). 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 of this study. The hydrodynamics for the two floater geometries can be found in Section 2.3.3.

The same sea state as in Figure 6.24 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 Figures 6.24a and 6.24b being set for each reinforcement learning 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.

In Figures 6.26a and 6.26b, the learning behaviours of the three control algorithms

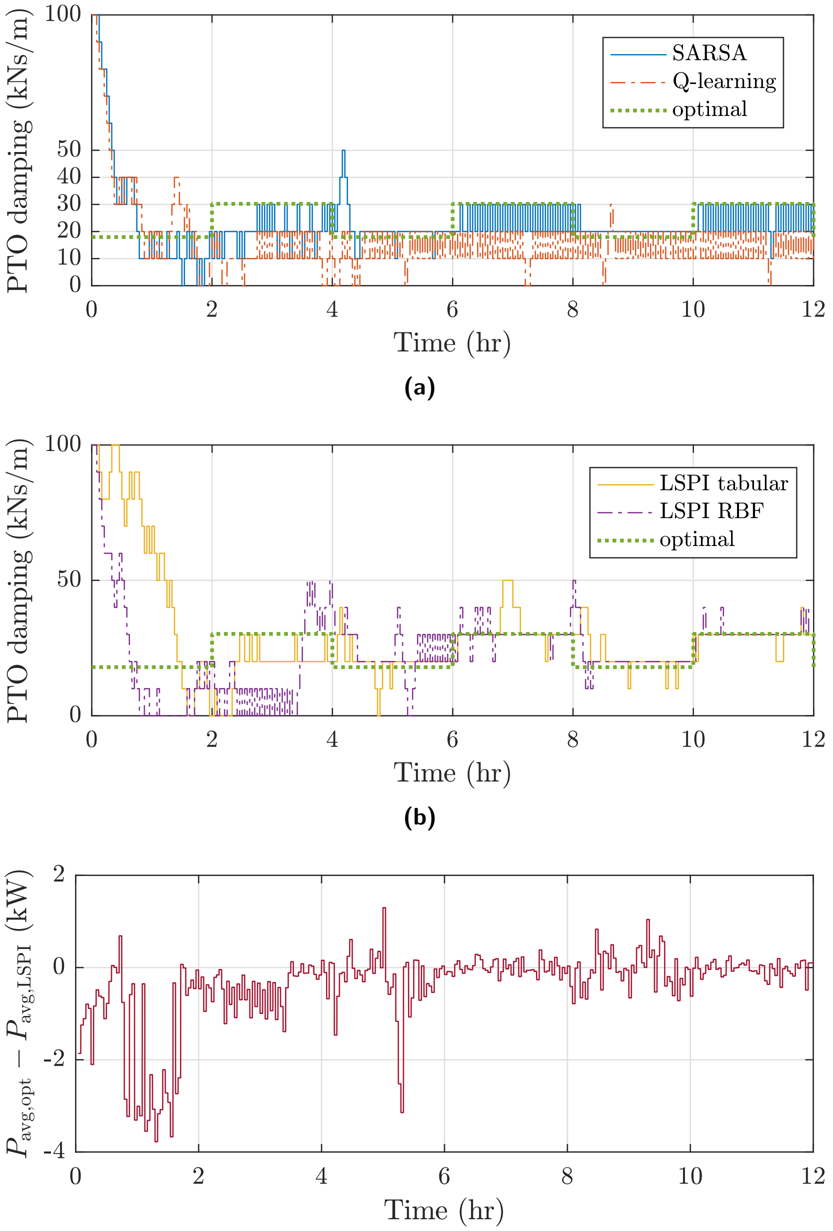


(a)

(b)

(c)

Figure 6.24: PTO damping coefficient selected by different reinforcement learning control strategies as compared with the optimal value in irregular waves with $H_{\text{s}} = 2$ m and $T_{\text{e}} = 6$ s and a JONSWAP spectrum starting from $B_{\text{PTO}} = 100$ kNs/m (a-b). (c) shows the difference in the mean generated power for the optimum ($P_{\text{avg,opt}}$) and the case of LSPI with RBFs ($P_{\text{avg,LSPI}}$).



(a)

(b)

(c)

Figure 6.25: PTO damping coefficient selected by different reinforcement learning 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 $B_\mathrm{PTO} = 100$ 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 ($P_{\mathrm{avg,LSPI}}$).

are shown. The difference in mean generated power between LSPI with RBFs and the optimal control setting is shown in Figure 6.26c.

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.

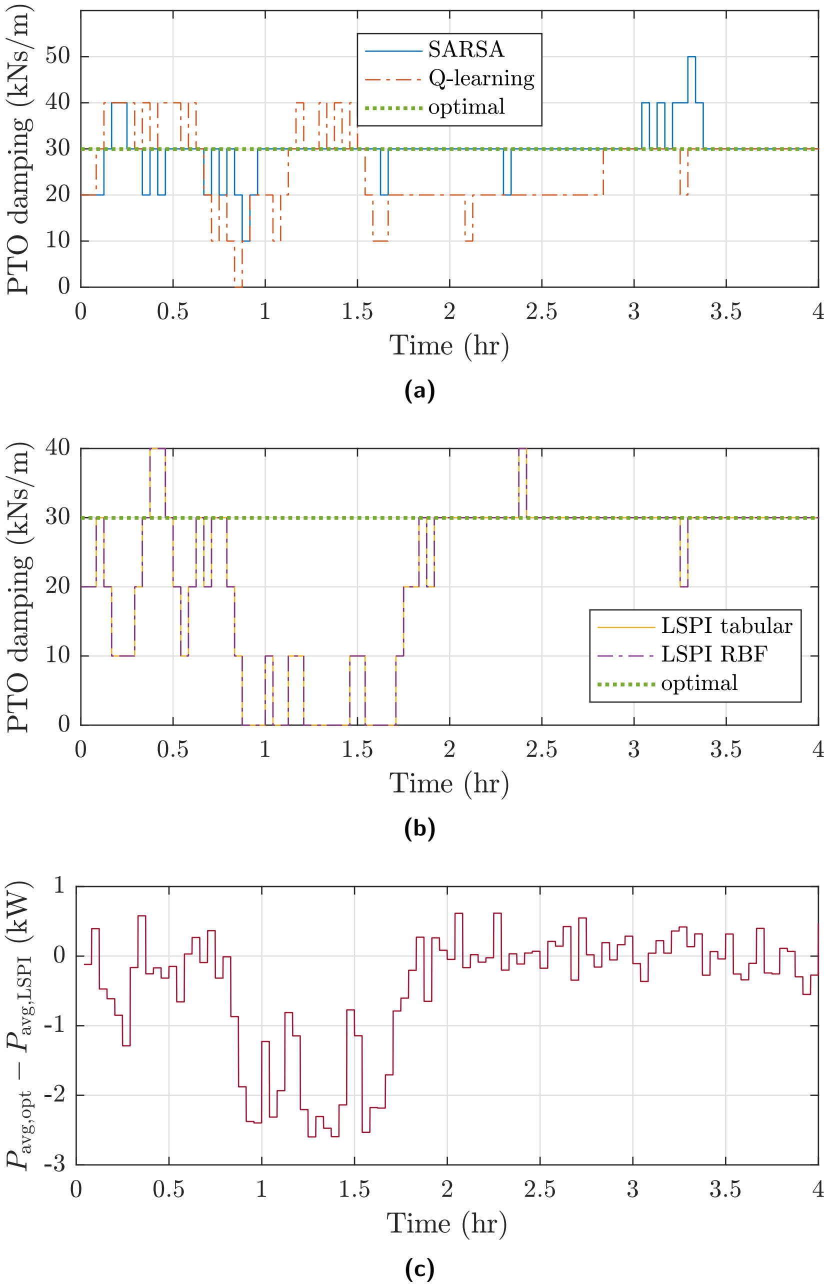
6.2.4.3 Discussion

Although in the previous case study no formal definition of convergence was attempted, here we consider the reinforcement learning 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 B_{PTO} within the one-hour period, which may be due to random actions being adopted.

6.2.4.3.1 Regular waves

In Figure 6.20, 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 B_{PTO} being very close to the starting PTO damping coefficient, thus requiring little exploration before finding the optimum. Conversely, Figure 6.21 represents a more challenging scenario for the 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 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 Figure 6.21b.

Comparing the behaviour of LSPI with tabular features and RBFs with $\delta_c = 10$ kNs/m and $\mu = 10$ kNs/m in Figures 6.20b, 6.21b and 6.24b, 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



(a)

(b)

(c)

Figure 6.26: PTO damping coefficient selected by different reinforcement learning 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 Figures 6.24a and 6.24b, respectively. (c) shows the difference in the mean generated power for the optimum ($P_{\mathrm{avg,opt}}$) and the case of LSPI with RBFs ($P_{\mathrm{avg,LSPI}}$).

number of kernels as discrete states are used, with the bandwidth spanning the space between discrete states. In Figure 6.22, decreasing the number of kernels was assumed to result in faster learning because the RBFs are expected to generalise the shape of the Q-function for unseen states and actions (Geramifard *et al.*, 2013). 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 B_{PTO} 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 reinforcement learning problems (as for instance in the examples in Chapter 4). Hence, the use of 5 or less RBFs incurs in an underfitting problem, i.e. using too coarse a model to fit the state-action value 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.

6.2.4.3.2 Irregular waves

Q-learning and Sarsa are similarly unable to converge towards the optimal policy in irregular waves as well, as shown in Figures 6.24 and 6.25. 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 Figure 6.25b. In particular, the learning time is lower than the 12 hours required by Q-learning for convergence in irregular waves in the previous case study, 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 (Lagoudakis and Parr, 2003). Furthermore, as shown in Figure 6.25b, 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 (Holthuijsen, 2007).

As the Seabased device is tested in the Skagerrak strait (Eriksson *et al.*, 2007), a JONSWAP spectrum is appropriate due to its bounded, shallow-water nature (Holthuijsen, 2007). However, a JONSWAP spectrum is a single-peaked spectrum with a relatively narrow frequency range (Holthuijsen, 2007). 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 reinforcement learning is expected to find the global optimum (Sutton and Barto, 1998), 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, reinforcement learning is proven to be able to adapt to changes in the dynamics of the WEC in Figure 6.26. Even though the reward function is initialized with the wrong values, reinforcement learning 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.

6.2.4.4 Summary

This case study proves that reinforcement learning is not affected by model non-linearities as expected. More importantly, LSPI is shown to achieve convergence towards the optimal passive control setting in less than 6 hours per sea state even in challenging conditions. This is a great improvement over the simple reinforcement learning control implementation of the previous section. At the same time, the controller is able to recognize a change in sea state and update the search for the PTO damping coefficient accordingly. In addition, the adaptive nature of reinforcement learning control is proven for all tested algorithms with a simple experiment.

The results with function approximation are more mixed, though. The expected improvement in performance has not materialised, possibly due to the small number of discrete values of the PTO damping coefficient employed. For a greater number of discrete values, greater benefits are expected with the use of function approximation, as shown in the studies on tidal energy not contained within this thesis or by the results of Wei *et al.* (2016) for the control of wind turbines. Nevertheless, the use of RBFs has been observed to stabilise the learning performance of the controller.

In both case studies, the influence of penalties for large displacements has not been studied in detail. This will be addressed in the next section on reactive, or impedance-matching, control, which results in a great increase in motion amplitude as well as energy absorption (Section 3.3). Similarly, the performance of reinforcement learning in broader banded spectra will be investigated.

6.3 Reactive control

In reactive, or impedance-matching, control, the controller can exert an additional stiffness term, i.e. the phase of the response can be controlled. As a result, the control problem no longer lies in the determination of an optimal damping coefficient per sea state, but rather the combination of a stiffness and a damping coefficient per sea state. In addition, since much greater power extraction and motion amplitude is expected due to the phase control, as shown in Section 3.3, penalties for large displacements will become more important than for resistive control.

The following sections will deal with the required changes in the reinforcement learning algorithm formulation as compared with resistive control.

6.3.1 State variables

The selected state space is as a result extended to:

$$\mathcal{S} = \left\{ s | s_{i,k,l} = \left(H_{\mathrm{s},i}, T_{\mathrm{e},k}, B_{\mathrm{PTO},l}, C_{\mathrm{PTO},n}\right), \begin{array}{c} i = 1:I, \\ k = 1:K, \\ l = 1:L \\ n = 1:N \end{array} \right\}, \tag{6.9}$$

where C_{PTO} is the PTO stiffness coefficient. The discretization of the state variables is performed as for resistive control. However, it is important to notice that some combinations of PTO damping and stiffness coefficients must not be selected, as these can result in an unstable behaviour. For point absorbers (and other WEC designs), a negative control stiffness coefficient may be required to achieve optimal performance in particular wave lengths (Wave Energy Scotland, 2016). This negative value can be achieved in practice by the electrical components of the PTO system. Nevertheless, it is fundamental to ensure that the negative PTO coefficient is never greater in magnitude than the positive hydrostatic stiffness coefficient, otherwise the system becomes unstable (Wave Energy Scotland, 2016).

For this reason, combinations with $B_{\mathrm{PTO}} = 0$ and $C_{\mathrm{PTO}} < 0$ are excluded from the search space. Preliminary simulations must be run for each particular design in order to identify the combinations of PTO damping and stiffness coefficients that would result in an unstable behaviour, so that they can be avoided.

6.3.2 Action Space

For reactive control, the action is a combination of increase, decrease, or not change the PTO damping and stiffness coefficients. This gives 9 possible actions as opposed to only 3 in the case of resistive control in Section 6.2. It has been preferred, however, to vary only one variable at a time in order to limit the action-state space thus decreasing the size of the state-action value function. This has a direct positive consequence on the overall learning time. The action space is thus given by:

$$\mathcal{A} = \{a|\,[(-\Delta B_{\text{PTO}},0),(0,-\Delta C_{\text{PTO}}),(0,0),(+\Delta B_{\text{PTO}},0),(0,+\Delta C_{\text{PTO}})]\}, \qquad (6.10)$$

where $\Delta B_{\text{PTO}} = B_{\text{PTO},l} - B_{\text{PTO},l-1}$ and $\Delta C_{\text{PTO}} = C_{\text{PTO},n} - C_{\text{PTO},n-1}$ are predefined step changes in the PTO damping and stiffness coefficients respectively.

The states corresponding to the minimum or maximum PTO damping and stiffness coefficients, i.e. $B_{\text{PTO},1}$, $B_{\text{PTO},L}$, $C_{\text{PTO},1}$ and $C_{\text{PTO},N}$, present a smaller action state to prevent the controller from exceeding the state space boundary. For instance, for $C_{\text{PTO},N}$, the action $+\Delta C_{\text{PTO}}$ is invalid.

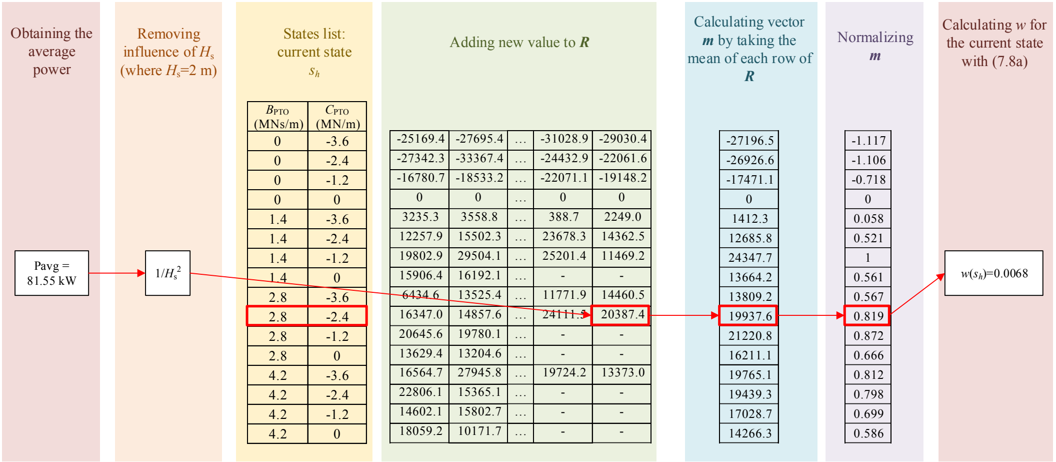
6.3.3 Reward function

The formulation of the reward function is the same as for resistive control in Section 6.2.0.3. However, in this case, due to the possible presence of negative power flows for a poor selection of PTO damping and stiffness coefficients, a separate normalization is performed for positive and negative power values, so that values are contained in the interval (-1,1) rather than (0,1) as in resistive control. The selection of an odd value of u is particularly important to prevent the rectification of negative power flows and the learning of a dangerous policy.

The reward function calculation process is summarized in Figure 6.27, where the data comes from the case study with the RM3 two-body point absorber introduced in Section 2.3.2. In the figure, $w(s_h)$ indicates the non-penalty term in (6.8) (i.e. when the displacement constraints are not active).

6.3.4 Algorithm

The reactive control algorithms based on reinforcement learning are very similar to those described in Section 6.2.1 for resistive control. The main difference is the increase of the state space to include the PTO stiffness coefficient. Furthermore, the power averaging is initialised only after a longer time interval than for resistive control, since the transient effects associated with a change in both PTO damping and stiffness coefficients are more significant.


Obtaining the average power
Pavg = 81.55 kW
Removing influence of H_s (where H_s=2 m)
$1/H_s^2$
States list: current state s_h
Adding new value to ***R***
Calculating vector ***m*** by taking the mean of each row of ***R***
Normalizing ***m***
Calculating w for the current state with (7.8a)
$w(s_h)$=0.0068

B_{PTO} (MNs/m)	C_{PTO} (MN/m)
0	-3.6
0	-2.4
0	-1.2
0	0
1.4	-3.6
1.4	-2.4
1.4	-1.2
1.4	0
2.8	-3.6
2.8	-2.4
2.8	-1.2
2.8	0
4.2	-3.6
4.2	-2.4
4.2	-1.2
4.2	0

-25169.4	-27695.4	…	-31028.9	-29030.4
-27342.3	-33367.4	…	-24432.9	-22061.6
-16780.7	-18533.2	…	-22071.1	-19148.2
0	0	…	0	0
3235.3	3558.8	…	388.7	2249.0
12257.9	15502.3	…	23678.3	14362.5
19802.9	29504.1	…	25201.4	11469.2
15906.4	16192.1	…	-	-
6434.6	13525.4	…	11771.9	14460.5
16347.0	14857.6	…	24111.5	20387.4
20645.6	19780.1	…	-	-
13629.4	13204.6	…	-	-
16564.7	27945.8	…	19724.2	13373.0
22806.1	15365.1	…	-	-
14602.1	15802.7	…	-	-
18059.2	10171.7	…	-	-

-27196.5	-1.117
-26926.6	-1.106
-17471.1	-0.718
0	0
1412.3	0.058
12685.8	0.521
24347.7	1
13664.2	0.561
13809.2	0.567
19937.6	0.819
21220.8	0.872
16211.1	0.666
19765.1	0.812
19439.3	0.798
17028.7	0.699
14266.3	0.586

Figure 6.27: Calculating the reward w (excluding penalties for large motions) at one step of the reinforcement learning algorithm in irregular waves. This corresponds to the last step in Figure 6.32 in Section 6.3.7.2.

Figure 6.28 shows the algorithm for the reactive control of a WEC using either Q-learning or Sarsa. If discrete states are used, the scheme should be built on top of Algorithms 3 and 2. When function approximation is employed, Algorithms 5 and 4 should be preferred instead. Conversely, Figure 6.29 displays the algorithm for the reactive control of a WEC using LSPI.

6.3.5 Displacement constraint handling controller

Although the penalty term in the reward function is effective in teaching the controller to avoid selecting combinations of the PTO coefficients that result in large motions, it does not prevent it from taking them. In fact, the agent needs to take those actions first in order to learn that they are bad. Including a safety factor in the displacement constraints reduces this risk considerably. Similarly, simulations can be used to pre-train the controller within a safe environment. When the control scheme is then applied to the actual device, the controller is expected to move only about the optimum point. Furthermore, the displacement constraints that must be met in order to prevent the application of a penalty should be set as soft constraints, i.e. with a magnitude smaller than the actual maximum allowable displacement. Nevertheless, these approaches do not remove the risk of exceeding the actual displacement constraints completely due to the random element in the ϵ-greedy exploration strategy in (4.9).

A solution is the implementation of a simple displacement-constraint-handling controller acting within reactive control. The controller operates directly on the PTO force in real time in order to try to keep the displacement at the PTO within safe limits.

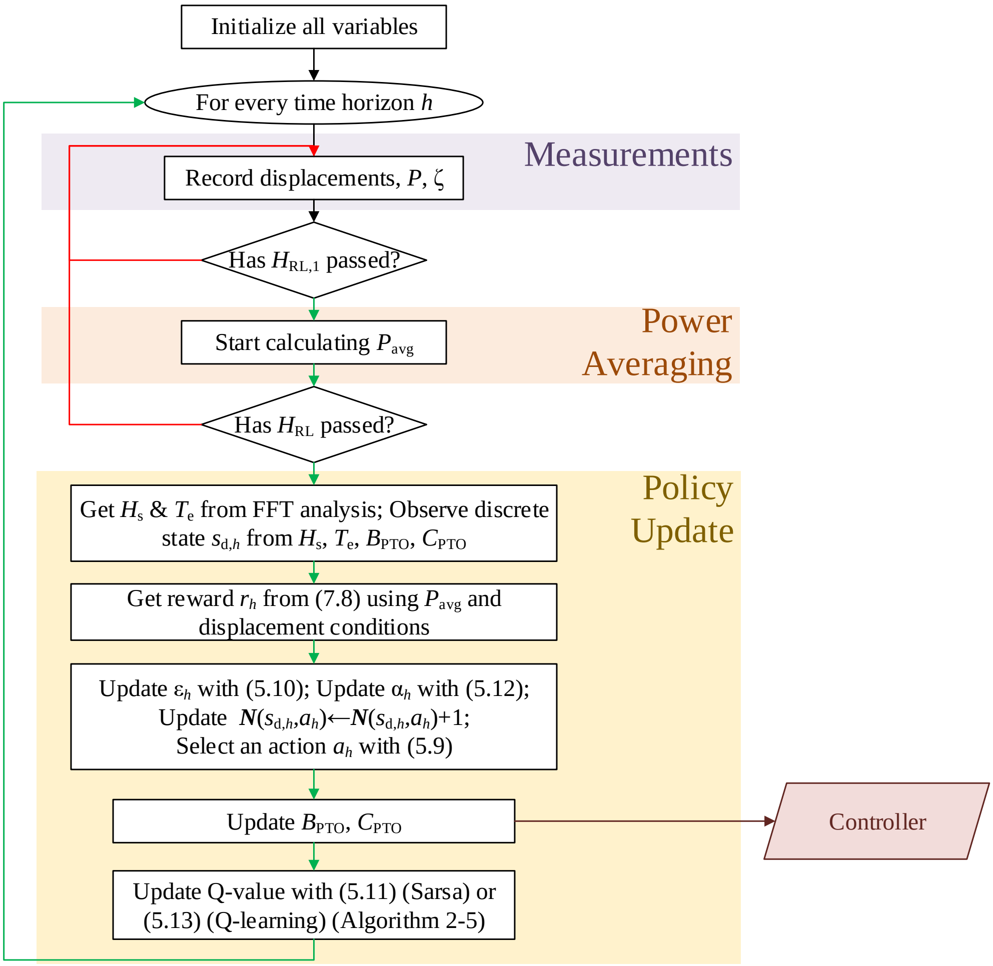

Initialize all variables
For every time horizon h
Measurements
Record displacements, P, ζ
Has $H_{\mathrm{RL},1}$ passed?
Power Averaging
Start calculating P_{avg}
Has H_{RL} passed?
Policy Update
Get H_{s} & T_{e} from FFT analysis; Observe discrete state $s_{\mathrm{d},h}$ from H_{s}, T_{e}, B_{PTO}, C_{PTO}
Get reward r_h from (7.8) using P_{avg} and displacement conditions
Update ε_h with (5.10); Update α_h with (5.12); Update $\boldsymbol{N}(s_{\mathrm{d},h},a_h) \leftarrow \boldsymbol{N}(s_{\mathrm{d},h},a_h)+1$; Select an action a_h with (5.9)
Update B_{PTO}, C_{PTO}
Controller
Update Q-value with (5.11) (Sarsa) or (5.13) (Q-learning) (Algorithm 2-5)

Figure 6.28: Flowchart of the Q-learning or Sarsa algorithm for the reactive control of a WEC.

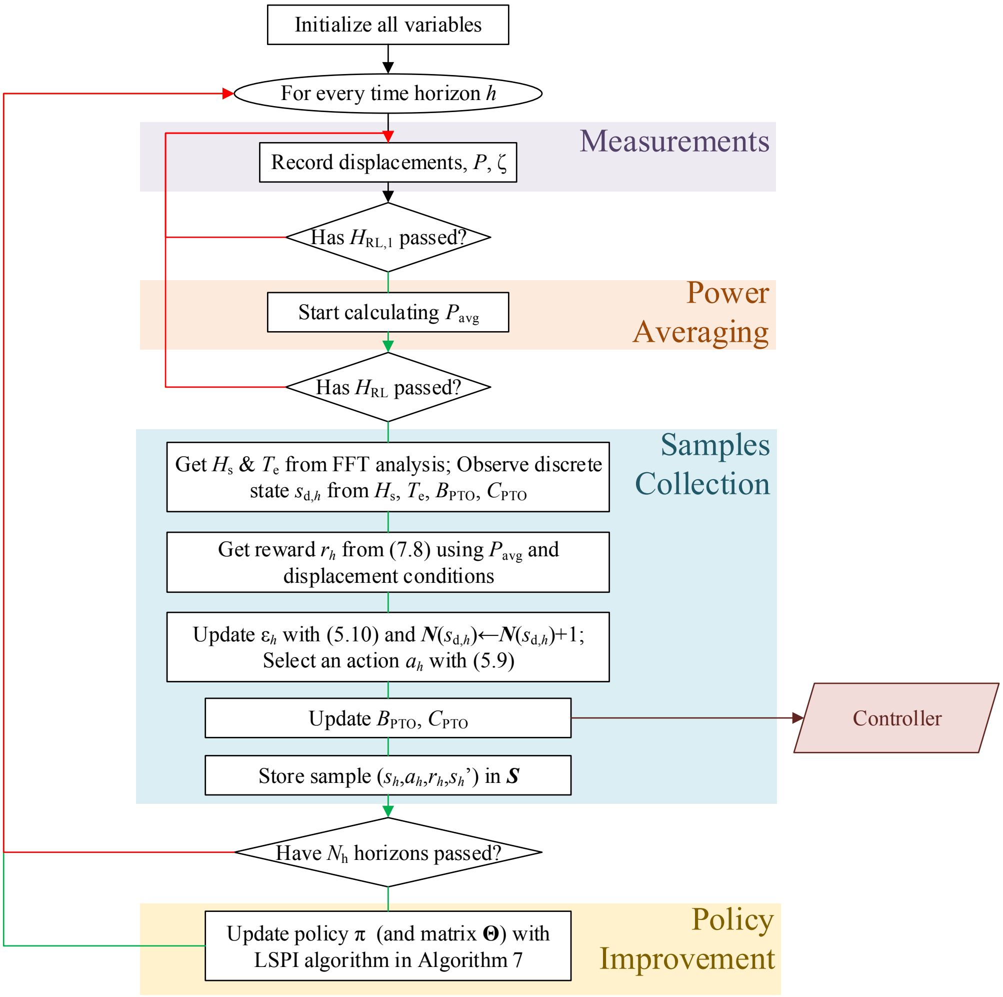

Initialize all variables
For every time horizon h
Measurements
Record displacements, P, ζ
Has $H_{\mathrm{RL},1}$ passed?
Power Averaging
Start calculating P_{avg}
Has H_{RL} passed?
Samples Collection
Get H_s & T_e from FFT analysis; Observe discrete state $s_{\mathrm{d},h}$ from H_s, T_e, B_{PTO}, C_{PTO}
Get reward r_h from (7.8) using P_{avg} and displacement conditions
Update ε_h with (5.10) and $N(s_{\mathrm{d},h})\leftarrow N(s_{\mathrm{d},h})+1$; Select an action a_h with (5.9)
Update B_{PTO}, C_{PTO}
Controller
Store sample (s_h,a_h,r_h,s_h') in $\boldsymbol{S}$
Have N_h horizons passed?
Policy Improvement
Update policy π (and matrix $\boldsymbol{\Theta}$) with LSPI algorithm in Algorithm 7

Figure 6.29: Flowchart of the LSPI algorithm for the reactive control of a WEC.

Whether this is feasible in practice or how it is implemented in reality is not treated within this work. A superior alternative would be the implementation of a controller with an optimization and prediction component similar to model-predictive control, but this task also goes beyond the scope of this work.

A very simple scheme is considered based on the magnitude of the instantaneous magnitude of the displacement and sign of the velocity of the float. Originally, if the magnitude is greater than a certain value that corresponds to a soft constraint, say $z_{\text{lim}} = 90\% z_{\text{Max}}$, and the sign of the velocity is either rising in a wave peak or decreasing in a wave trough, the applied PTO force is changed in sign:

$$f_{\text{PTO,rt}}(t) = \begin{cases} -f_{\text{PTO}}(t) & \text{if } z > z_{\text{lim}} \ \& \ \dot{z} > 0 \text{ or } z < -z_{\text{lim}} \ \& \ \dot{z} < 0 \ , \quad (6.11a) \\ f_{\text{PTO}}(t) & \text{otherwise,} \quad (6.11b) \end{cases}$$

where f_{PTO} is obtained as usual for reactive control as described in Section 3.3. However, it is clear that the switch in sign relies on the assumption of the stiffness term dominating the PTO force. If this is not the case, then a switch in sign is strongly undesirable, since the damping effect would be lost and in fact inverted. As a result, an improved simple controller is proposed in the form:

$$f_{\text{PTO,rt}}(t) = \begin{cases} 0 & \text{if } z > z_{\text{lim}} \ \& \ \dot{z} > 0 \text{ or } z < -z_{\text{lim}} \ \& \ \dot{z} < 0, \quad (6.12a) \\ f_{\text{PTO}}(t) & \text{otherwise.} \quad (6.12b) \end{cases}$$

It should be noted that z_{Max} is obtained from the sensitivity analysis of the displacement on the PTO coefficients in each sea state using simulations and it already includes a safety factor, as aforementioned.

This extra controller will be studied only in the second case study, in Section 6.3.8.

6.3.6 Case studies

The proposed reactive control for WECs based on reinforcement learning has been tested on two numerical benchmark cases:

- the reference model 3 (RM3) point absorber that comprises of a float and a reaction plate, first introduced in Section 2.3.2;
- the linear point absorber first introduced in Section 2.3.1.

Only the Q-learning algorithm is applied to the first case study, which represents the proof of concept. Although the device presents two bodies, the PTO has only one degree of freedom so that the problem is simplified considerably. Furthermore, this case study presents a Bretschenider spectrum, typical of the West Coast of the United States (Neary *et al.*, 2014). The second case study, done on a simpler model in a JONSWAP

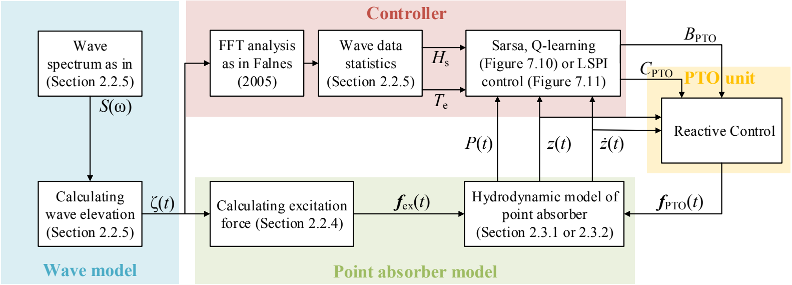

Controller
Wave spectrum as in (Section 2.2.5)
S(ω)
FFT analysis as in Falnes (2005)
Wave data statistics (Section 2.2.5)
H_s
T_e
Sarsa, Q-learning (Figure 7.10) or LSPI control (Figure 7.11)
B_PTO
C_PTO
PTO unit
Reactive Control
P(t)
z(t)
ż(t)
Calculating wave elevation (Section 2.2.5)
ζ(t)
Calculating excitation force (Section 2.2.4)
f_ex(t)
Hydrodynamic model of point absorber (Section 2.3.1 or 2.3.2)
f_PTO(t)
Wave model
Point absorber model

Figure 6.30: Work-flow diagram of the program used to simulate the point absorber with reinforcement-learning-based reactive control.

spectrum, is used mainly to address the problem of displacement constraint abidance of the proposed reinforcement learning strategies.

In both case studies, the simulation system shown in Figure 6.30 has been employed.

6.3.7 Case study: RM3 two-body point absorber

The maximum PTO force that can be exerted due to the generator rating has been assumed to be $F_{\mathrm{Max}} = 1$ MN, while the magnitude of the maximum displacement at the PTO has been limited to $x_{\mathrm{PTO,Max}} = 5$ m. The PTO efficiency is set as $\eta = 80\%$.

For simplicity, the PTO damping coefficient is assumed to range from 0 to 4.2 MNs/m in steps of 1.4 MNs/m, so that $L = 4$. Similarly, the PTO stiffness coefficient is taken to range from -3.6 MN/m to 0 MN/m in steps of 1.2 MN/m, so that $N = 4$. These values have been selected as they fully enclose the optimal coefficients for the analysed sea states. As a result of the choice of PTO damping and stiffness coefficients, 16 reinforcement learning states are used when a single sea state, as given by H_{s} and T_{e}, is considered. Nevertheless, for a more realistic implementation a finer resolution and a wider range are expected.

In the Q-learning algorithm, the discount factor and the initial exploration and learning rates have been set to $\gamma = 0.95$, $\epsilon_0 = 0.6$ and $\alpha_0 = 0.4$, respectively. The parameters for the decay of the exploration and learning rates have been set to $N_\epsilon = 25$ and $N_\alpha = 5$, respectively. The power of the reward function is set to $u = 25$ and the penalty for large displacements to $p = 2$. In addition, in both regular and irregular waves, the time horizon duration has been set to $H_{\mathrm{RL}} = 20T_{\mathrm{e}}$. Due to the longer transient period after the change in PTO coefficients, the power averaging process is started only after a period of $H_{\mathrm{RL}} = 8T_{\mathrm{e}}$ from the start of the period.

In both regular and irregular waves, a single sea state (i.e. $I = K = 1$) is analysed for simplicity, since the reinforcement learning control has been shown to learn in multiple sea states for resistive control previously (Section 6.2).

6.3.7.1 Results in regular waves

A single sea state, i.e. $I = J = 1$, has been analysed in regular waves, with unit amplitude and a wave period of 8 s. The wave trace lasts 8 hours after a preliminary initialization, which is discarded. Figures 6.31a and 6.31b compare the curves of the PTO damping and stiffness coefficients respectively with time as selected by the Q-learning algorithm against the optimal values. The difference in the corresponding mean absorbed power and the optimal mean generated power of 260.5 kW (obtained as per Section 3.3 with a 20-minute-long wave trace) can be seen in Figure 6.31c.

6.3.7.2 Results in irregular waves

Similarly, a single sea state, with a significant wave height of 2 m and a peak wave period of 9.25 s is considered in irregular waves with a Bretschneider spectrum as a proof of concept. From the Fast Fourier Transform analysis, the energy wave period for the generated wave trace is 8 s. The wave time series is 8 hours long. As per the regular waves case, $I = J = 1$ so that the reinforcement learning problem reduces to 16 states.

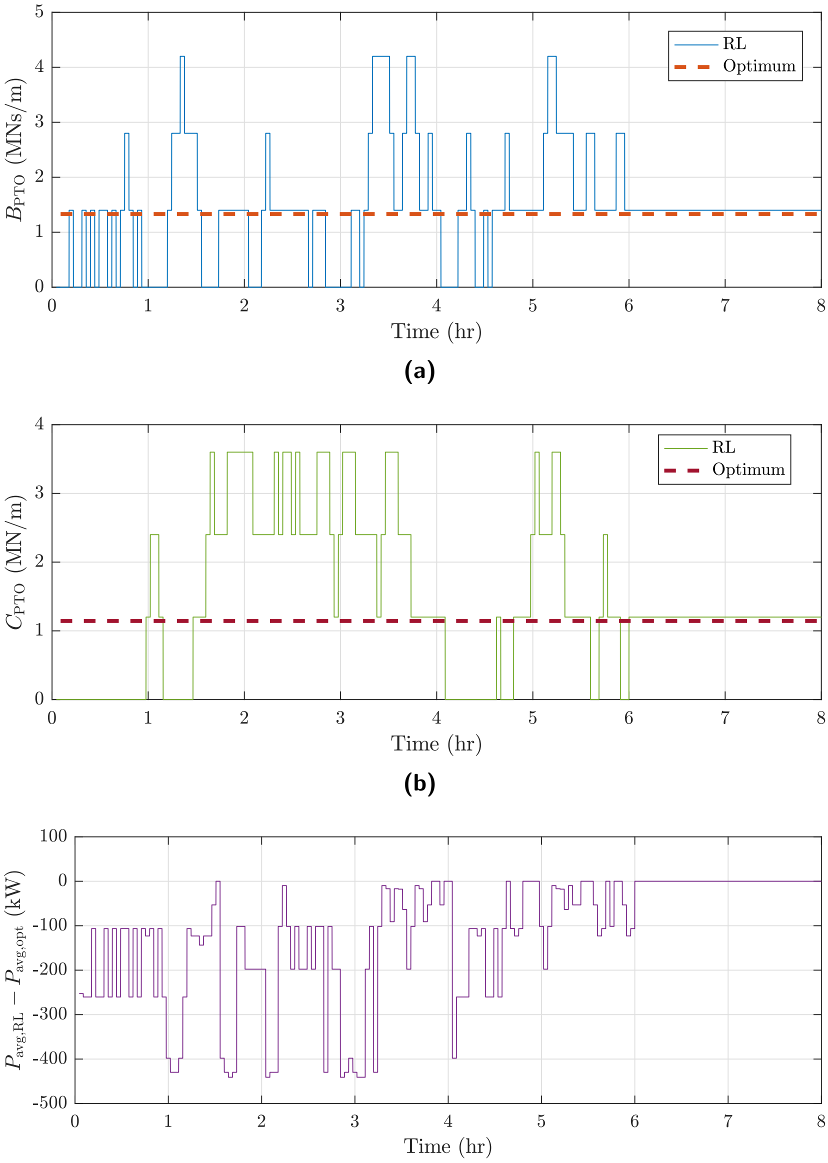
In Figures 6.32a and 6.32b, it is possible to see the PTO damping and stiffness coefficients respectively adopted by the reinforcement learning control scheme as compared with the optimal values in this sea state. Figure 6.32c shows the difference in the corresponding mean absorbed power, with the mean generated power obtained by using the optimal coefficients being 90.582 kW.

6.3.7.3 Discussion

6.3.7.3.1 Regular waves

As is clear from Figure 6.31, in regular waves the Q-learning algorithm learns the optimal PTO coefficients in approximately six hours from a random start ($\boldsymbol{Q} = \mathbf{0}$). This is almost double the time required by the control scheme for resistive control in Section 6.2.2 mainly due to the longer time horizon employed: $20T_\mathrm{e}$ as opposed to $10T_\mathrm{z}$ or $10T_\mathrm{e}$, with the energy wave period being typically greater than the zero-crossing mean wave period. In fact, a shorter time horizon may be used considering the deterministic nature of regular waves. Additionally, the convergence time is strongly dependent on the number of discrete B_PTO and C_PTO values employed, with only 16 states currently being used.

In Figure 6.31, it is also interesting to notice the random initial behaviour of the controller due to the selected exploration strategy, which enables the agent to visit

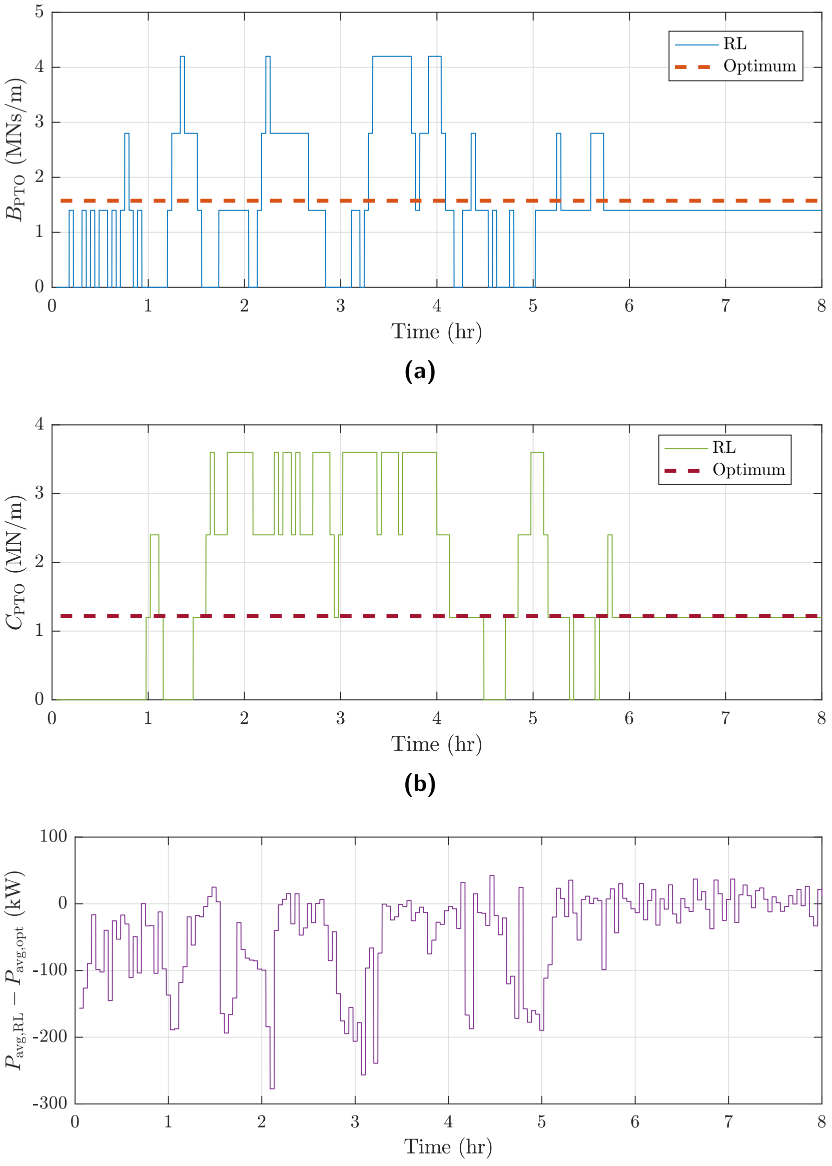


(a)

(b)

(c)

Figure 6.31: Time variation of the PTO damping (a) and stiffness (b) coefficients chosen by the Q-learning control as compared with the respective optimal values in regular waves of unit amplitude and a wave period of 8 s. (c) shows the difference between the corresponding mean generated power and the optimal mean generated power.



(a)

(b)

(c)

Figure 6.32: Time variation of the PTO damping (a) and stiffness (b) coefficients chosen by the reinforcement learning control as compared with the respective optimal values in irregular waves with $H_{\text{s}} = 2$ m and $T_{\text{e}} = 8$ s. (c) shows the difference between the corresponding mean generated power and the optimal mean generated power.

most states. As the learning progresses, the exploration rate tends to zero and the algorithm chooses the optimal, exploitative actions.

In order to meet the requirements of the linear wave theory assumption of the hydrodynamic model, a short wave height has been chosen. As a result, the prescribed maximum PTO displacement is never exceeded. Hence, the penalty term in (6.8) is not applied. If it were, the controller would be expected to select a higher PTO damping coefficient, like for resistive control. Conversely, a PTO stiffness coefficient with a smaller, if not zero, magnitude is forecast, as the controller tries to move away from resonance. On the other hand, the force reaches the saturation limit even in this mild sea state. However, a bang-bang behaviour similar to the one in Section 6.2.2 is not observed with reactive control.

6.3.7.3.2 Irregular waves

From Figure 6.32, it is evident that the developed statistical reward function is effective in ensuring convergence in irregular waves as well, despite their stochastic nature. Furthermore, since the same horizon time length is employed as per the regular waves run, the learning time is no greater than Q-learning for resistive control as in Section 6.2.2. Nevertheless, the challenge that irregular waves pose to the convergence of the correct action selection can be understood by comparing Figures 6.31c and 6.32c, where the much more oscillatory nature of the mean absorbed power in irregular waves is clear.

A typical sea state has a duration that ranges between 30 minutes and 6 hours (Holthuijsen, 2007). Hence, even though the learning time is smaller than in Section 6.2.2 despite the larger number of states, convergence is still unlikely to be achieved before there is a variation in the significant wave height and energy wave period. However, as shown in Section 6.2.2 for irregular waves with multiple sea states, the Q-learning algorithm applied to the control of WECs is able to pick up the learning process from where it left off the last time it encountered a particular sea state. This represents the main advantage of reinforcement learning over traditional optimization algorithms, which would be unable to identify whether a change in the cost function is due to a change in the PTO damping or stiffness coefficients or due to noise in the wave energy.

In a realistic application, a finer grid of B_{PTO} and C_{PTO} values would be desired in order to deal with a large range of sea states. Nevertheless, this may increase the learning time excessively and will be studied in the next section. The state-action value function is expected to be pre-initialized through numerical simulations in order to prevent selecting PTO settings that result in excessive motions in energetic sea states, which could be a real problem with reactive control.

Finally, it is important to understand that reinforcement learning is proposed as a method to remove the dependence of existing WEC control strategies from hydrodynamic models. Therefore, the overall controller performance is only as good as the control scheme itself, with reactive control representing a significant improvement over resistive control treated in Section 6.2.

6.3.7.4 Summary

In this case study, the Q-learning algorithm has been applied to the reactive control of a two-body point absorber. This application basically represents a generalization of the resistive control scheme, with a corresponding greater search space. This case study further supports the applicability of reinforcement learning control to different WEC technologies.

In the next case study, the performance of reinforcement learning control under displacement constraints conditions is analysed.

6.3.8 Case study: linear point absorber

In this case study, the simple linear point absorber constrained to motions in heave first introduced in Section 2.3.1 is considered. The overall PTO efficiency has been set to $\eta = 0.75$ and the force constraint to $F_{\mathrm{Max}} = 0.5$ MN. Only LSPI is applied to the reactive control of the point absorber.

Simulations have been run in both regular and irregular waves. In regular waves, the wave height and period have been set to $H = 2$ m and $T = 8$ s, respectively. As a single sea state is considered, $I = K = 1$ and $J = LN$. In irregular waves, a JONSWAP wave spectrum has been adopted. In order to demonstrate the ability of reinforcement learning to switch between different sea states as for resistive control, two alternating sea states are considered. The first sea state presents $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 7$ s, while the latter $H_{\mathrm{s}} = 2$ m and $T_{\mathrm{e}} = 8$ s. The two sea states alternate every 2 hours, which is a realistic duration for a sea state, with typical values ranging from 0.5 to 6 hours (Holthuijsen, 2007). Therefore, $I = 1$, $K = 2$ for the simulations in irregular waves.

Using the LSPI algorithm with discrete features, the PTO damping and stiffness coefficients are each discretized with 7 values, ranging from 0 to 300 kNs/m and -300 to 0 kN/m with steps of 50 kNs/m and 50 kN/m, respectively. This relatively fine discretization seems realistic for practical applications. In irregular waves, only 4 values are used for the PTO damping coefficient (hence, in steps of 100 kNs/m) in order to speed up convergence, since the mean generated power is more affected by the PTO stiffness coefficient. As a result, the total number of states for each sea state is 49 and 56 in regular and irregular waves, respectively.

Additionally, an optimization is run to find the optimal coefficients for each sea states in the analysis as described in Section 3.3. This has been used in the presentation of the results as a benchmark for the reinforcement learning solution.

6.3.8.1 Results with no displacement constraints

Initially, the simulations are run with the float displacement limit set at $z_{\text{Max}} = 5$ m, which is never exceeded in either regular or irregular waves for the selected sea states.

6.3.8.1.1 Regular waves

A wave trace lasting 6 hours is generated. The time variation of the PTO damping and stiffness coefficients selected by the LSPI algorithm can be seen in Figures 6.33a and 6.33b, respectively. Figure 6.33c shows the difference between the generated power and the power generated using the optimal coefficients, used as a benchmark.

6.3.8.1.2 Irregular waves

A 24-hour long time series is employed, with the sea states alternating every 2 hours. The reinforcement learning control action and the corresponding generated power can be seen in Fig. 6.34. It can be seen that the first sea state ($H_{\text{s}} = 2$ m and $T_{\text{e}} = 7$ s) is run for an extra hour to show that the wiggle in the B_{PTO} value just before $t = 22$ hr is due to random actions being selected by the ϵ-greedy exploration strategy.

6.3.8.2 Results with displacement constraints active

In order to assess the efficacy of the penalty formulation, the displacement constraint has been lowered to ±2 m, which is lower than the amplitude of the response achieved with the optimal PTO setting. This is preferred over an increase in the energy content of waves because a linear model is used for the hydrodynamics, whose validity is void for large motions. Only regular waves have been analysed for this study, considering a 10-hour-long wave trace.

The control action selection of reinforcement learning can be seen in Figure 6.35 as compared with the optimal coefficients. These are calculated by modifying the cost function of the optimization to return a power value of 0 if the displacement constraint is exceeded. The difference in the generated power between the reinforcement learning response and the optimal solution is shown in Figure 6.35c. Note that for the combinations $B_{\text{PTO}} = 200$ kNs/m and $C_{\text{PTO}} = -250$ kN/m, and $B_{\text{PTO}} = 200$ kNs/m and $C_{\text{PTO}} = -300$ kN/m the constraint is exceeded.

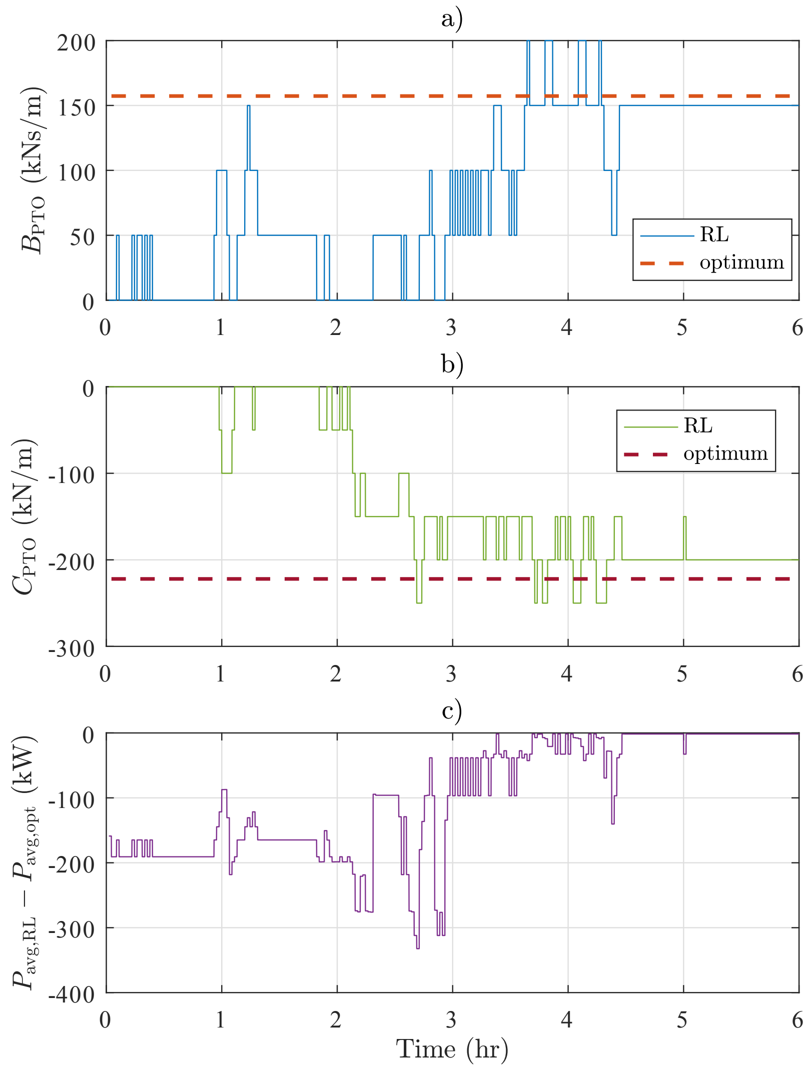


Figure 6.33: Selection of the PTO damping (a) and stiffness (b) coefficients by the reinforcement learning control as compared with the respective optimal values in regular waves with $H = 2$ m and $T = 8$ s and a maximum allowable displacement of 5 m. The difference in the corresponding mean generated power can be seen in (c).

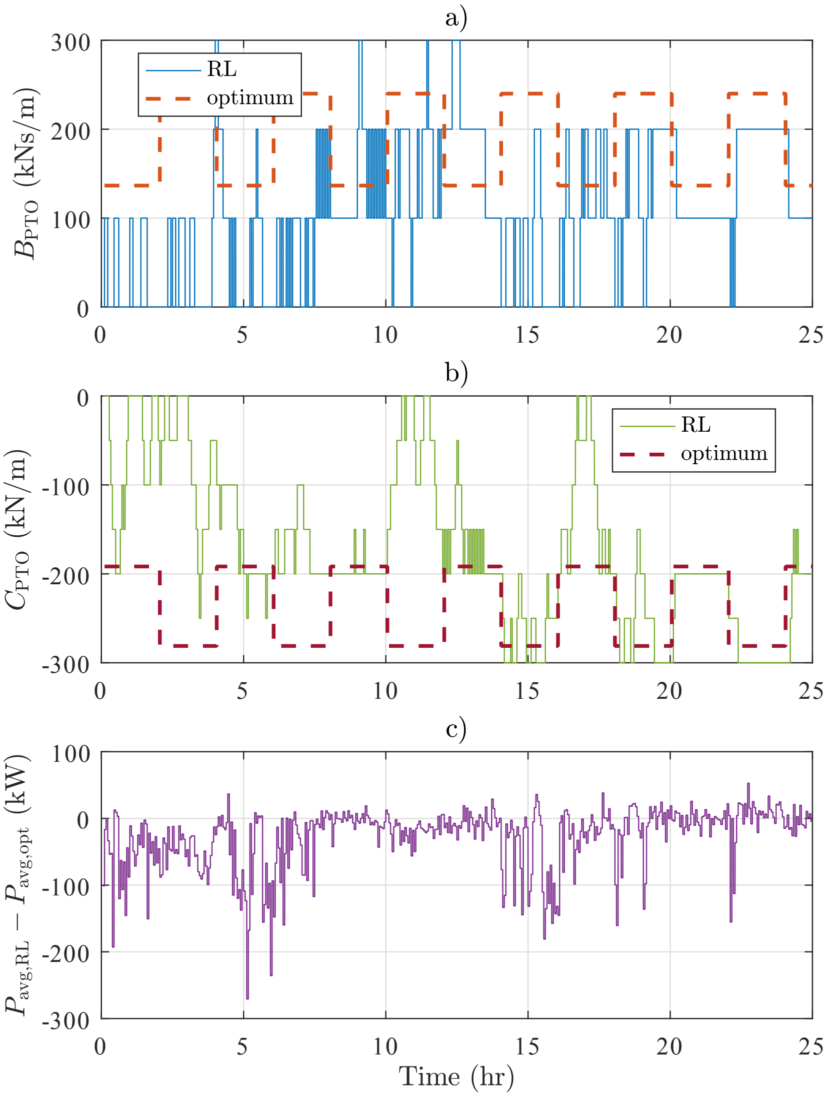


Figure 6.34: Time variation of the PTO damping (a) and stiffness (b) coefficients chosen by the reinforcement learning control as compared with the respective optimal values in irregular waves with two, alternating sea states. (c) shows the difference between the corresponding and the optimal mean generated power.

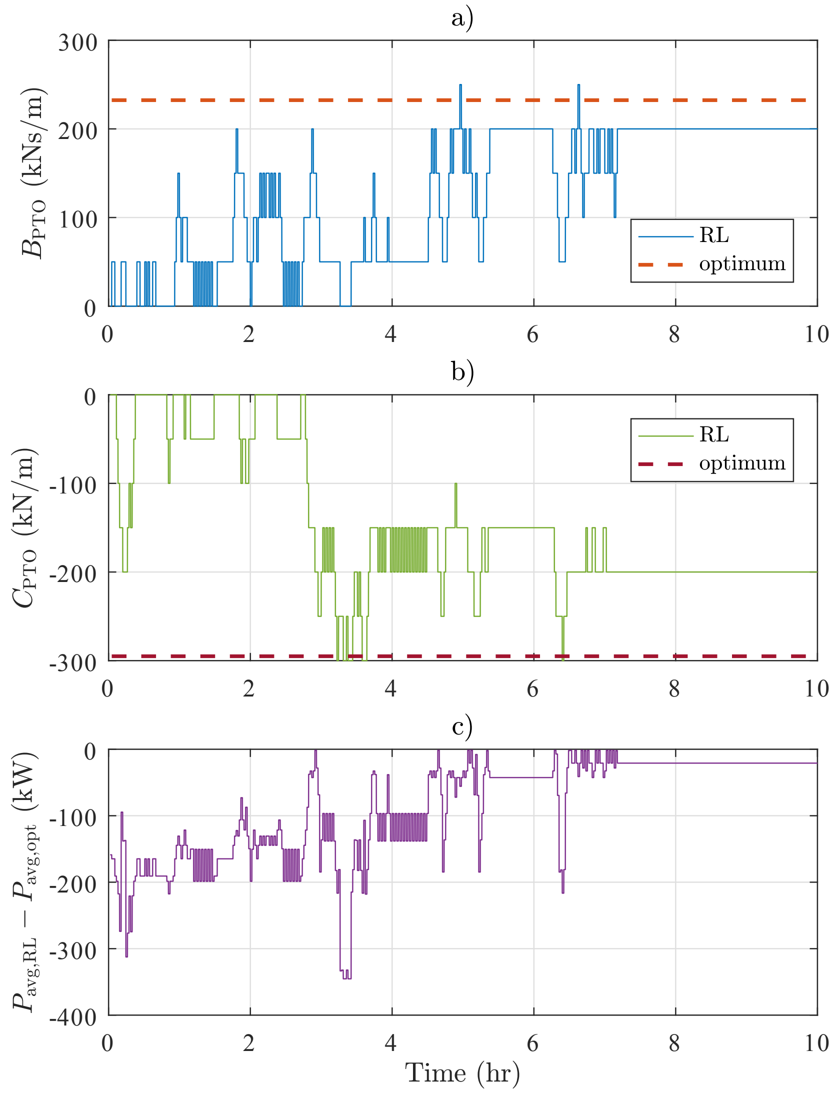


Figure 6.35: Time variation of the PTO damping (a) and stiffness (b) coefficients chosen by the reinforcement learning control as compared with the respective optimal values in regular waves with $H = 2$ m and $T = 8$ s and a maximum allowable displacement of 2 m. (c) shows the difference between the corresponding and the optimal mean generated power.

6.3.8.3 Results for the real-time controller for soft displacement constraints

The performance of the proposed low-level controller is assessed with simulations in both regular and irregular waves. In particular, the coefficients learned by LSPI in the unconstrained runs are employed in conjunction with a soft constraint $z_{\text{lim}} = 0.9 z_{\text{Max}} = 1.8$ m. For clarity, the constrained response of the point absorber is plotted against the unconstrained response.

6.3.8.3.1 Regular waves

In regular waves with $H = 2$ m and $T = 8$ s, the PTO coefficients are set to $B_{\text{PTO}} = 150$ kNs/m and $C_{\text{PTO}} = -200$ kN/m. A 120-s-long time series is sufficient to get a fully-developed response, as shown in Figure 6.36 for both constrained (continuous line) and unconstrained (dotted line) cases. In particular, Figure 6.36a presents the float displacement and the wave elevation ζ, Figure 6.36b the float velocity, Figure 6.36c the PTO displacement, and Figure 6.36d the instantaneous and mean generated power.

6.3.8.3.2 Irregular waves

Considering only one sea state with $H_{\text{s}} = 2$ m and $T_{\text{e}} = 8$ s and a JONSWAP spectrum, the PTO coefficients are set to $B_{\text{PTO}} = 200$ kNs/m and $C_{\text{PTO}} = -300$ kN/m. Figure 6.37 shows the response of the point absorber for a portion of the wave trace with a higher energy content. In particular, Figure 6.36a presents the float displacement and the wave elevation ζ, Figure 6.36b the float velocity, Figure 6.36c the PTO displacement, and Figure 6.36d the instantaneous and mean generated power.

6.3.8.4 Discussion

6.3.8.4.1 Reinforcement learning analysis

By looking at Figures 6.33 and 6.34, LSPI with discrete states and actions is found to learn the optimal coefficients in both regular and irregular waves when no displacement constraints are active. However, the selected large number of states results in very slow learning time. This is particularly evident in irregular waves, where up to 12 hours are required for convergence per sea state. Nevertheless, this figure shows that reinforcement learning is able to recognize the change in sea state and pick up learning from where it left off the last time the controller encountered those wave conditions. This is fundamental for a practical implementation of LSPI control of a WEC. The learning time strongly depends on the number of states and thus on the discretization of H_{s}, T_{e}, B_{PTO} and C_{PTO}. For this reason, alternative machine learning schemes that provide a

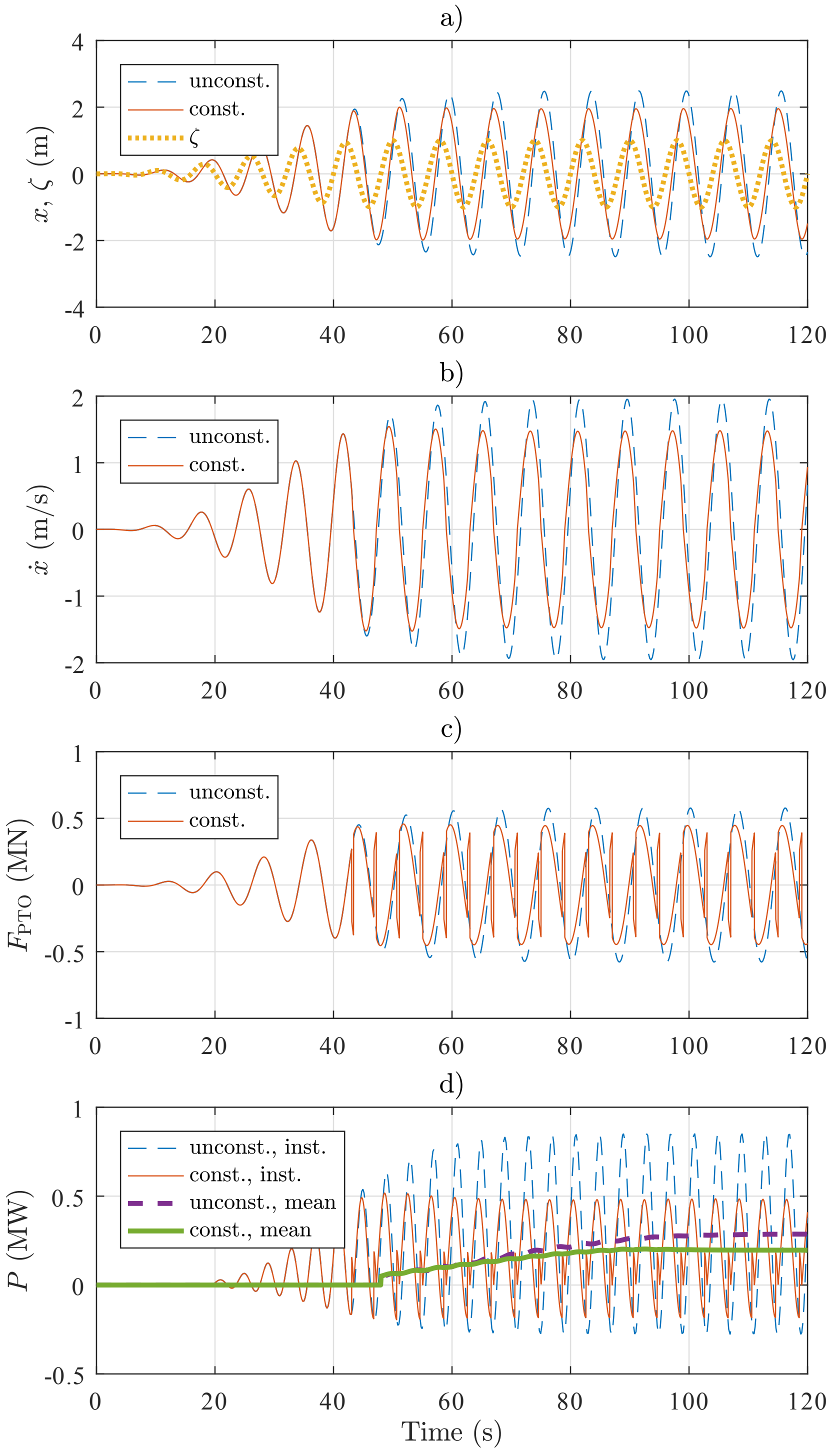


Figure 6.36: Response of the device in both constrained and unconstrained conditions in regular waves with $H = 2$ m and $T = 8$ s, including plots of the displacement, velocity, PTO force and generated power.

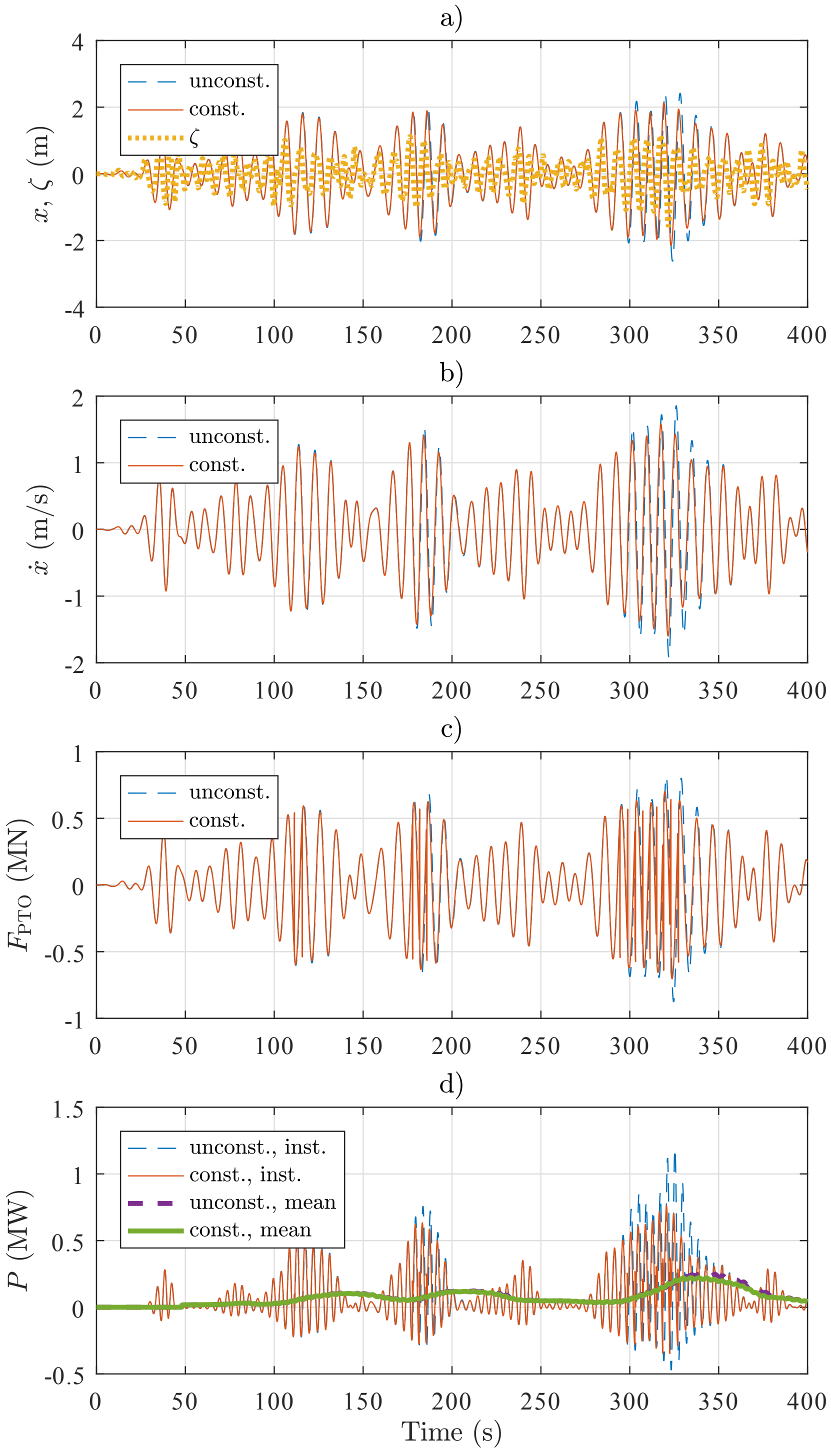


Figure 6.37: Response of the device in both constrained and unconstrained conditions in irregular waves with $H_\mathrm{s} = 2$ m and $T_\mathrm{e} = 8$ s, including plots of the displacement, velocity, PTO force and generated power.

continuous regression, such as artificial neural networks, which have been investigated in the previous chapter, may be superior.

From Figure 6.35, when the displacement constraints are active, LSPI seems to be unable to converge towards the optimal coefficients. In fact, under closer scrutiny it *does* learn the optimal coefficients, since the penalty term is active for both combinations $B_{\text{PTO}} = 200$ kNs/m and $C_{\text{PTO}} = -250$ kN/m, and $B_{\text{PTO}} = 200$ kNs/m and $C_{\text{PTO}} = -300$ kN/m. Hence, the response of reinforcement learning is affected by the inability to select a PTO damping coefficient closer to 220 kNs/m, which results in a power loss of about 30 kW as compared with the optimal response, as can be seen in Figure 6.35c. The sensitivity of the mean generated power on the PTO coefficients with reactive control is another indicator that a continuous optimization method, such as the one based on neural networks in Chapter 5, is superior for this application.

Another worrying feature that can be seen in Figure 6.35 is the selection of combinations of the PTO coefficients that result in the exceedance of the displacement constraint during the exploration stage of the reinforcement learning algorithms. Although the algorithm does learn to avoid these states because of the penalty term, the fact that they are encountered at all would result in failure in practice. A solution would be to pre-train reinforcement learning with simulations, so that it learns to avoid some extreme actions. Once applied to the actual device, the controller would then correct the rewards it obtained from the simulations from those observed in reality. Nevertheless, this does not completely remove the possibility of selecting catastrophic actions. Hence, the proposed low-level controller is likely to be required as a fall-back option in any case.

6.3.8.4.2 Real-time controller for soft displacement constraints

As is clear from Figure 6.36a, the proposed real-time controller is able to limit the float displacement within ±2 m in regular waves despite the use of soft rather than hard constraints. The action of the controller is evident in Figure 6.36c from the comparison between the constrained and unconstrained cases. Whether such a response is practically feasible is another problem that will need addressing. Furthermore, as expected the application of the constraints results in a drop in mean generated power in Figure 6.36d.

In irregular waves, a particularly challenging situation has been analysed, with energetic waves relative to the selected sea state. As can be seen in Figure 6.37, the soft constraints are not able to prevent the float from exceeding the limits, despite the magnitude of the displacement is only just greater than 2 m. This is because of the steepness of the response, which means the effect of the PTO force of opposite sign comes too late.

This shows that the selected controller is too simplistic, and more accurate studies are required in order to prevent exceedance of displacement constraints in realistic wave conditions. In particular, the value of 90% use to determine the soft constraints will need to be adjusted based on the device dynamics and the sea states of interest.

6.3.8.5 Summary

In this case study, LSPI with discrete features has been applied to the reactive control of a simple point absorber. A fine discretization has been used for the PTO damping and stiffness coefficients. As a result, the algorithm requires 12 hours to learn the optimal policy in each sea state.

The treatment of displacement constraints has been analysed in detail. Although the penalty term in the reward function ensures that the algorithm learns to avoid large displacements, these need to be experienced first, which may cause to damages to or failure of the WEC. Therefore, it is possibly best to train the controller with simulations first, so that during actual operations the controller will explore only an area near the optimum. An alternative approach with a simple, lower-level control has also been addressed, but this is shown to be overly simplistic and case specific.

6.4 Chapter summary

In this chapter, reinforcement learning has been successfully applied to the control of WECs.

To start with, Monte-Carlo methods have been used in a simple implementation of declutching control. Although there is a performance gain over resistive control, the approach is too simplistic and inferior to declutching control with optimal command theory.

Subsequently, a practical control for the resistive control of WECs is developed based on reinforcement learning. The scheme is then generalized to reactive control. The controller finds the optimal PTO coefficients in each sea state for the maximization of energy absorption. At the same time, realistic force constraints are considered as well as penalties for large displacements. Three different algorithms are analysed: Q-learning, Sarsa and LSPI. Similarly, three different WEC models are used to assess the convergence properties of the algorithms for different levels of abstraction.

From the case studies analysed within this section, the following observations have been made:

- LSPI provides superior performance to Q-learning and Sarsa, with a learning time of up to 6 hours per sea state with resistive control. This increases to 12 hours with reactive control due to the wider search space.
- The controller is able to recognize changes in sea states. In particular, whenever it enters a particular sea state, the agent will pick up learning from where it left off the last time that sea state was encountered.
- The reinforcement learning schemes are not affected by system non-linearities due to their model-free nature.
- The controller is able to adapt to changes in the system dynamics. These may be either abrupt, as in the case of non-critical sub-system failures, or slower, for instance due to marine growth and components ageing.
- Function approximation has not been found to be particularly beneficial for resistive control due to the small search space. However, it improves the stability of the learning behaviour.
- The proposed control strategy accounts for realistic saturation constraints on the PTO force. If the force saturates, the controller finds the optimal policy for the maximization of energy absorption that applies to the new scenario. In some cases, this seems to converge towards a bang-bang type of control.
- The penalty term on large displacements is effective, since the controller learns to avoid those actions. However, in order to do so, the agent needs to experience those conditions, which may lead to failure.
- An additional, lower-level controller has been developed to try to address the problem of displacement constraint exceedance. Nevertheless, it has been found to be overly simplistic and case-specific.

Therefore, the following conclusions may be made:

- A practical implementation of resistive control with reinforcement learning is feasible.
- Conversely, the larger state-action space associated with reactive control means that learning can be very slow. Function approximation may alleviate this problem by speeding up convergence.
- In these studies, sea states have been assumed to be stationary for the duration of a time horizon (approximately 3-8 minutes). However, this is not true in practice, as shown by the necessary application of statistical techniques for the development of the reward function. Energy is transported in packets by wave groups. So, realistically in order to maximize performance the controller should optimize the controller parameters for each wave group.
- Similarly, to ensure the maximization of energy absorption, the reduction of loads and the abidance by displacement constraints, it is necessary to include prediction within the algorithm, as done in the previous chapter. If the controller operates on

a wave-group-by-wave-group basis, the prediction becomes feasible if a network of buoys is installed at a sufficient distance in front of the WEC.

- The controller should be coupled with a higher-order controller that would shut down the system to survival mode if an extreme storm is forecast, when the PTO action is predicted to be poor due to force saturation.
- Simulations should be used to pre-train the reinforcement learning behaviour in energetic sea states, so that the controller learns to avoid the selection of parameters that may lead to failure.
- The exploration rate will need to be periodically reset, even though possibly by a small amount, so that the controller is able to adapt to any possible changes in the system dynamics in the meantime.

Chapter 7

Conclusions

The main contribution of this work has been the development of algorithms based on reinforcement learning for the passive and active control of wave energy converters (WECs). Focus was given to a practical implementation that is realistic as inspired by the state-of-the-art control strategies employed by the wave energy industry. Therefore, rather than being optimized in real-time, the controller parameters are selected for a particular range of wave conditions (known as sea states), which are assumed stationary over a time period lasting tens of wave cycles. The sea state is indicated by the wave height and wave period in regular waves, and the significant wave height and energy wave period in irregular waves. The control parameters are represented by a damping coefficient (passive or resistive control) or a damping and a stiffness coefficient (active or reactive control). Within the reinforcement learning framework, at the start of each time period the controller is in a particular state, as given by the combination of the sea state and control parameters. It then selects an action which is a change (or no change) in control parameters and updates the control force. At the end of the control signal, the controller receives a reward, which is a function of the mean generated power in the time interval. If soft displacement constraints are exceeded during the time period, a penalty is returned instead. At the start of the new time interval, the controller is in a new state and the cycle is repeated. Through reinforcement learning, the controller learns an optimal behaviour with time for the maximization of the total reward.

In this thesis, we have shown that reinforcement learning successfully converges towards the optimal coefficients in simulation studies, as obtained from an optimization. In particular, the controller is able to recognize changes in sea states and update the control parameters accordingly. This is achieved without reliance on models of the system dynamics. For this reason, the control strategy is unaffected by system non-linearities, as shown in this thesis by different case studies using simulations with models of increasing complexity. In addition, reinforcement learning is proven to adapt to sudden changes in system dynamics. This could correspond to non-critical subsystem failures or slower effects due to marine growth. The adaptive characteristics of reinforcement learning are fundamental in achieving an increase in the levellised of energy produced by WECs through an increase in capacity factor and system performance.

Detailed studies have been conducted to assess the performance of different reinforcement learning algorithms. In contrast with results typical of the robotics industry, function approximation has not been found to be particularly beneficial, mainly due to the small number of states employed. The most promising strategy that has been analysed, least-squares policy iteration, is shown to find the optimal control parameters within 6 and 12 hours in irregular waves for passive and active control, respectively. These figures are considered realistic for a practical implementation considering the expected lifetime of a WEC (approximately 20 years). Furthermore, although sea states typically last up to 6 hours in fact, reinforcement learning is proven to recognize changes in sea state and pick up learning from where the controller left off the last time that particular sea state was encountered. In addition, reinforcement learning is shown to adapt its response if realistic saturation limits are reached for the force exerted by the power take-off system (PTO). The control scheme can also learn to avoid parameters that would result in the exceedance of displacement constraints in energetic waves, which may lead to damages to or failure of the device. However, in order to learn this behaviour, the controller needs to experience the sea states and actions first, which is undesirable. Finally, thanks to its model-free nature, reinforcement learning control has been shown to converge towards the optimal control parameters even when non-linear effects are present, e.g. due to the PTO unit, and to adapt to changes in the system dynamics.

An alternative approach based on artificial neural networks has also been proposed for the reactive control of WECs. This technique is also based on the assumption of stationary sea conditions for a time period lasting tens of wave cycles. A neural network is used to produce the non-linear mapping between mean generated power and maximum displacement amplitude with significant wave height, wave energy period and control parameters in each time interval. This can be considered to be a type of system identification. A global optimization scheme based on computational performance (through exploitation of parallel processing) is suggested for the determination of the control parameters at the start of each time horizon based on the predicted significant wave height and energy wave period. An exploratory technique is used to investigate the search space, with the focus shifting towards the selection of the expected optimal action as the number of sample points increases. Nevertheless, this approach may result in the selection of control parameters that result in highly negative power flows or an exceedance of the displacement constraints due to the randomness of the exploration strategy. In irregular waves, the controller is shown to learn faster the optimal control parameters than reinforcement learning due to the continuous nature of its features.

A final merit of this thesis is the detailed description of reinforcement learning and its introduction to the wave energy industry. Although only simple, theoretical (i.e. not treating the control design in detail, including its interactions with the hardware and monitoring capabilities) implementations have been proposed in this work, the thesis shows the potential of this method so that it may be further developed by the wave energy industry.

7.1 Limitations of the proposed methods

The proposed methods present attractive advantages, but do not come without limitations.

In this thesis, the control strategies developed based on reinforcement learning and artificial neural networks rely on the assumption of stationary wave conditions over a fixed number of wave cycles. Although it is possible to statistically recognize stationary sea states, greater power absorption is expected from a control strategy that can adapt to the wave excitation force in real time. More importantly, a real-time approach would provide a more robust approach for dealing with constraints on the displacement, velocity, force and power flow at the PTO units or joints. The proposed reinforcement learning and neural network methods result in a conservative behaviour due to the assumption of stationary wave conditions. Indeed, in irregular waves, the displacement constraint may be exceeded only for a very short period of time, e.g. over one wave cycle (in the order of 7-15 s), over the whole time horizon (in the order of 200-400 s). Yet, a penalty is still returned, which will teach the controller to avoid the controller parameters that caused the constraint exceedance in that particular sea state. Therefore, it is clear that a superior approach must be developed for the handling of the displacement constraints, which is likely to be performed in real time.

Another issue is represented by the fact that the proposed reinforcement learning and neural network methods can account for penalties on the displacement only after experiencing the situations that should be avoided. For this reason, the controller should be trained with simulation and experimental data before being applied to a prototype WEC. Additionally, in the event of unforeseen circumstances that cause the device to experience untested conditions, linear theory should be employed to provide a first guess to prevent completely random behaviour.

A further limitation is represented by the lack of a prediction component in the proposed reinforcement learning schemes. This enables a fully model-free approach. Nevertheless, from the literature it is clear that the wave elevation forecasting is very important for the maximization of energy absorption and the meeting of realistic constraints. A prediction component could be included in a reinforcement learning scheme, although this would require a statistical model (hence, using dynamic programming). Conversely, the proposed neural networks strategy employs information on the predicted wave conditions. In practice, for stationary wave conditions, sea-state data can be obtained from meteorological organizations.

Finally, reinforcement learning is most suitable when there is a discrete number of possible actions. Hence, bang-bang types of control would be more appropriate than the developed strategies for resistive and reactive control of WECs.

7.2 Future work

Based on the drawn conclusions and identified limitations of the proposed methods, the following recommendation for future work are made.

The energy content in waves is subdivided into packets, known as wave groups. With the use of a network of buoys around a WEC (or, more economically in the future, a WEC array), it would be possible to determine incoming wave groups. Hence, the controller could be designed to adapt the control parameters based on the predicted incoming wave groups. Modifying the proposed approaches to the treatment of wave groups would make the control schemes more responsive, with a predicted increase in performance. The prediction of incoming wave groups would also be feasible with the use of wave buoys, as opposed to the forecast of the wave conditions over future tens of wave cycles. Nevertheless, transient effects associated with a change in parameters would become more relevant, which requires the development of effective solutions for this problem.

In addition, the method combining neural networks as a modelling tool and an optimization function should be carried forward for reactive control due to its superior performance over reinforcement learning for the determination of the optimal control parameters with reactive control. Nevertheless, the neural network must be pre-trained using simulations (including non-linear models) and tank test experiments to prevent the selection of random actions that may result in failure.

Conversely, the discrete nature of reinforcement learning indicates that its use would be best for a bang-bang type of control. Hence, a more realistic implementation of declutching control should be investigated. Latching control also represents an ideal candidate, with associated higher energy extraction than for declutching control. In fact, latching and declutching control could be considered within the same framework to further increase energy absorption, as suggested by Clément and Babarit (2012) with optimal command theory instead of reinforcement learning. Serious thought should be dedicated to constraint abidance, though. Furthermore, the use of function approximation is considered to be necessary if the state space is based on the WEC motions.

Finally, once methods are finalised for the control of an individual WEC, the schemes should be extended to the treatment of arrays of WECs. Using the proposed methods, this generalization should be straightforward. The achievement of economies of scale is fundamental in lowering the levellised cost of energy generated by WECs.

7.3 Recommendations

Although the proposed strategies are applicable to any devices, the methods have been validated with linear models (including only some non-linear effects in the PTO). Hence, it is recommended to assess their performance of these methods with fully non-linear models, such as virtual wave tanks (in CFD), or experimentally in wave tanks.

At the moment, the proposed methods are conservative in the handling of displacement constraints due to their time-averaging nature. Hence, it would be interesting to develop a framework that is able to deal with constraints in real-time to improve the performance of the proposed strategies. This could be either a reformulation of the reinforcement learning problem or simply the inclusion of a controller that deals with the constraints.

Finally, this thesis has developed the methods from a theoretical perspective. Therefore, the algorithms have been started from random initial conditions to show their convergence capabilities. Nevertheless, for real-applications, the results coming from linear wave theory should be used as an initial guess to prevent undesired behaviour and possible damage to the WEC. In particular, existing control strategies, such as model predictive control, may be used to control the WEC, with machine learning schemes, such as neural networks, providing the adaptive component by correcting the explicit model of the scheme with the observed data.

7.4 Concluding remarks

The development of an effective control strategy is fundamental in achieving a necessary increase in the levellised cost of energy extracted from ocean waves. Not only are the benefits associated with a rise in power absorption, but also in a reduction of loads and consequent damages to WECs. Machine learning approaches are shown to provide successful control strategies, which do not rely on models of the system dynamics for the determination of the optimal control actions. As a result, the system can adapt to both varying wave conditions and also changes in the system dynamics, e.g. due to marine growth or non-critical subsytem failures. Their potential should be further explored by the wave energy industry.

Appendix A

Neural Networks

In this chapter, we discuss some neural networks strategies that will later be applied to the control of WECs. First of all, we discuss briefly the development of neural networks and the main architectures. Then, we move on to the treatment of the simple feed-forward neural networks, which will be used in this thesis. In particular, we show their use as function approximation tools and some of the approaches used in their training. Finally, we present a test case to show the performance of the described neural network type.

A.1 Background

Artificial neural networks (ANNs) are a class of supervised learning algorithms that are inspired from nature. In supervised learning schemes, the machine learns the mapping between input and output data that is provided by a supervisor, i.e. the user, in a training data set. The machine learning algorithms will then generalised the knowledge to unseen situations and predict the output corresponding to the desired input data.

Like the biological brain which comprises of many interconnected neural cells, ANNs are made up of a network of interconnected functional units, known as nodes. By combining multiple neurons in a number of layers, so that the output of the neurons in one layer becomes the input to the neurons in the next layer, ANNs can be used to fit non-linear functions with a large number of input values. However, ANNs are no longer limited to use as function approximators. In fact, they have recently become ubiquitous in a number of disciplines and applications, with some of the most famous examples being the deep learning solutions to image and speech recognition in computer science (LeCun *et al.*, 2015). Indeed, deep learning can be considered to be the training of ANNs with a very large number of layers.

Due to their importance, there is a very large number of publications and even journals, such as Neural Networks and Neurocomputing, on ANNs. In this thesis, we rely mainly on the reviews on deep learning by Schmidhuber (2015) and LeCun *et al.* (2015), with the former providing an overview of the history of the development of ANN techniques up to deep learning and the latter focusing mainly on the recent advances, with a focus on image, word and speech recognition. The treatment of the historical development of neural networks goes beyond the scope of this work and the reader is referred to those works for greater information. Nørgaard *et al.* (2003) focus on the application of ANNs to modelling and control. In addition, the book by Hagan *et al.* (1996) supplies a practical description of the basic ANN methods, which will be employed in this work. Indeed, the power of deep learning goes beyond the requirements of the control of WECs for the time being, with more interest being possible once arrays of WECs are deployed.

Since the first study on neural networks by McCulloch and Pitts (1943), different network architectures and learning strategies have been developed. The first architecture to be proposed was the perceptron (Rosenblatt, 1958), which was soon found to have a limited set of applications. Since then multi-layer perceptrons, feed-forward, recurrent and convolutional neural networks have been investigated (LeCun *et al.*, 2015). Feed-forward neural networks are one of the simplest architectures, but they have been employed effectively for non-linear function approximation. Convolutional neural networks, which rely on local connections, shared weights, pooling and many layers, have become the standard in image and speech recognition applications (LeCun *et al.*, 2015). In recurrent ANNs, the output of some neurons is fed back to previous layers of the network (Mandic and Chambers, 2001). As a result, they are more sensitive to noise and may become unstable. Nevertheless, they have been successfully applied to system identification in WECs by Valério *et al.* (2008) and Giorgi *et al.* (2016b). The employed recurrent ANN types are neural network auto-regressive with exogenous inputs (Leontaritis and Billings, 1985a,b) and locally recurrent network (Elman, 1990). However, in this work ANNs are used mainly to produce a non-linear mapping between specified input and output variables. As a result, the simpler feed-forward architecture is employed.

Learning methods for ANNs are summarized in Atiya (1991). Over the years, many strategies have been developed including associative, performance and competitive learning, which originate from the work by Hebb (1949), Widrow and Hoff (1960) and Rosenblatt (1958) respectively. In this work, we consider backpropagation (Rumelhart *et al.*, 1986), which is a performance learning scheme. This algorithm updates the weights of the neural network by propagating the error signal backwards through the layers of the network. This scheme is behind the success of deep learning and it is nowadays widely adopted for the training of most ANN applications (LeCun *et al.*, 2015).

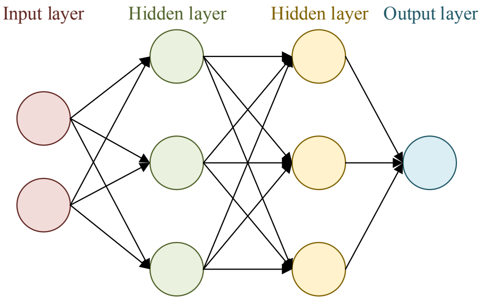


Figure A.1: Graphic representation of a simple feed-foward ANN.

A.2 Feed-forward neural networks

As aforementioned, feed-forward networks are a simple type of ANNs where the signal is propagated forward. A graphic representation of an example can be seen in Figure A.1. The network consists of neurons arranged in a number of layers. The signal is propagated forward from the input to the output layer along the arrows, which correspond to synapses in the biological brain. Similarly, the signal along each arrow and into each neuron is different from the neighbouring ones, as would the case in the real brain based on the activation of the specific connections for a particular task. In a mathematical framework, this is achieved by allocating a weight for each arrow and an *activation function* at each node. Training occurs by tweaking the weights so as to obtain a match between the prediction of the ANN and the provided output data based for the same input data.

The model of each neuron is described in the next section, including the most common activation functions. Then, a mathematical model for the whole network is obtained in Section A.2.2.

A.2.1 Neuron model

Figure A.2 shows the diagram of a single neuron or peceptron from a mathematical perspective. The input to the neuron is represented by the vector $\boldsymbol{x}$, which comprises of two element in this example. The signal from each input node n is then multiplied by a weight w_n before being summed to the other signals (Hagan *et al.*, 1996). This can be expressed in matrix form as

$$d = \boldsymbol{w}^T \boldsymbol{x}. \tag{A.1}$$

It is clear that to prevent imbalances and numerical instabilities, each input value should be normalized prior to being fed to the neuron. The signal d is then passed through an

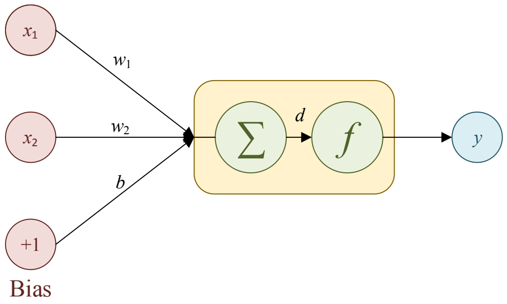


Figure A.2: Diagram of a single neuron or perceptron.

activation function f to yield the output of the neuron (Hagan *et al.*, 1996)

$$y = f(d). \tag{A.2}$$

An additional type of node exists known as *bias*, which is characterized by a fixed value of 1 (Hagan *et al.*, 1996). The associated signal presents a weight b. From a mathematical perspective, the bias provides the offset of the function fit. It can be omitted from a particular layer if desired. In case the input layer to the neuron presents a bias term as for instance in Figure A.2, then (A.1) is modified to

$$d = \boldsymbol{w}^T \boldsymbol{x} + b. \tag{A.3}$$

Different activation functions have been proposed over the years, and these will be discussed next.

A.2.1.1 Activation functions

The activation function is known as such because it shows the amount by which the neuron is activated for given input signals. In general, non-linear functions are employed with values ranging from 0 to 1 or -1 to 1 to prevent numerical errors and excessive imbalances. One of the simplest activation functions is the hard limit activation function (Hagan *et al.*, 1996), which is obtained as

$$y = \begin{cases} 1 & \text{if } d > 0, \qquad \text{(A.4a)} \\ -a & \text{otherwise,} \qquad \text{(A.4b)} \end{cases}$$

where $a = 0$ or $a = 1$ for the asymmetrical and symmetrical hard limit function, respectively. An alternative simple activation function, unconstrained to $[0, 1]$ or $[-1, 1]$, is the linear function

$$y = d. \tag{A.5}$$

From this function, it is also possible to obtain the rectified linear function

$$y = \max(0, d), \tag{A.6}$$

which is analogous to half-wave rectification (Hagan *et al.*, 1996). Additional activation functions are the asymmetric and symmetric linear saturation functions

$$y = \begin{cases} 1 & \text{if } d > 1, & \text{(A.7a)} \\ d & -a \leq d \leq 1, & \text{(A.7b)} \\ -a & \text{if } d < -a, & \text{(A.7c)} \end{cases}$$

where $a = 0$ or $a = 1$ for the asymmetrical and symmetrical saturation linear function, respectively.

Nevertheless, these functions (with the exception of the linear function) present at least one point of discontinuity at which the derivative is undefined. In the backpropagation process, as discussed in Section A.2.4, the derivative of the activation function is required for the training of the ANN. For this reason, two smooth functions with a shape similar to the linear saturation function have been the study of most applications until recently (Hagan *et al.*, 1996; LeCun *et al.*, 2015): the asymmetric standard logistic sigmoid and the symmetric hyperbolic tangent functions.

The standard logistic sigmoid (or simply sigmoid) function is expressed as Hagan *et al.* (1996)

$$y = \frac{1}{1 + e^{-d}}. \tag{A.8}$$

In fact, it corresponds to the derivative of the smooth approximation to the rectified linear function: the soft-plus function

$$y = \ln\left(1 + e^{d}\right). \tag{A.9}$$

The derivative of the sigmoid function itself can be calculated as

$$\dot{f}(d) = \frac{e^{d}}{\left(1 + e^{d}\right)^{2}} = f(d)\left(1 - f(d)\right), \tag{A.10}$$

where f indicates the sigmoid function.

Similarly, the hyperbolic tangent is given by

$$y = \tanh(d) = \frac{1 - e^{-2d}}{1 + e^{-2d}}, \tag{A.11}$$

with derivative

$$\dot{f}(d) = \frac{e^{2d} - 1}{e^{2d} + 1} = 1 - \tanh^{2}(d) = 1 - f(d)^{2}, \tag{A.12}$$

where f indicates here the hyperbolic tangent function.

The linear function presents a derivative with value 1 for all values of d, while the rectified linear function a value of 0 for $d < 0$ and 1 for $d > 0$, with a discontinuity at $d = 0$.

The functions described in (A.4), (A.5), (A.6), (A.7), (A.8) and (A.11) are shown graphically in Figure A.3. Recently, the rectified linear function in (A.6) has become the mainstay of deep learning applications, as it enables faster learning for applications with a large number of layers (LeCun *et al.*, 2015). However, in this work, only a few layers will be employed. For this reason, the smoother hyperbolic tangent in (A.11) has been used as the activation function in the input and hidden layers. Conversely, the output layer presents a linear activation function (A.5), since ANNs are applied to a regression rather than a classification task (Hagan *et al.*, 1996). This means that ANNs are employed as a function approximation method rather than to classify items in a limited number of categories.

A.2.2 Network model

The feed-forward neural network in Figure A.1 is expressed with the mathematical model of the perceptron of Figure A.2 in Figure A.4. In this example, the ANN presents an input layer with two values, two hidden layers with three nodes each and an output layer with a single variable. Each layer is described by an index l, with $l = 1$ and $l = 4$ corresponding to the input and output layers, respectively. Each node presents an output ${o^l}_n$, with l and n indicating the indices of the layer and the position of the node within the layer, respectively. Similarly, each node presents the activation variable ${d^l}_n$. Therefore, each layer can be described by the vectors $\boldsymbol{d}$ and $\boldsymbol{o}$. Furthermore, each layer (except for the input layer) is characterized by the activation function f^l. The input layer is represented by the vector of parameters $\boldsymbol{x}$ while the output layer by $\boldsymbol{y}$ for each sample. Each arrow, or synapsis, has an associated weight ${W^l}_{i,j}$ that is characterized by the indices l, i and j, which represent the layer that precedes it, the node that is the starting point in layer l and the node that the arrow points to in layer $l+1$. Hence, each sets of arrows can be represented by the weight matrix $\boldsymbol{W}^l$. In addition, all layers other than the output layer have an associated bias term, with corresponding vector $\boldsymbol{b}^l$.

In order to assess the performance of ANNs and determine the system geometry (i.e. the number of layers and neurons per layer), after rearranging the points randomly, the original data set is usually divided into three subsets: the training, validation and test sets (Ng, 2016). The training data set, which typically comprises 50% to 80% of the original data points, is employed for the training of the ANN. For this reason, it is used to determine the end conditions for training after performance convergence is achieved. The validation data set, which is usually made of 10%-25% of the original data

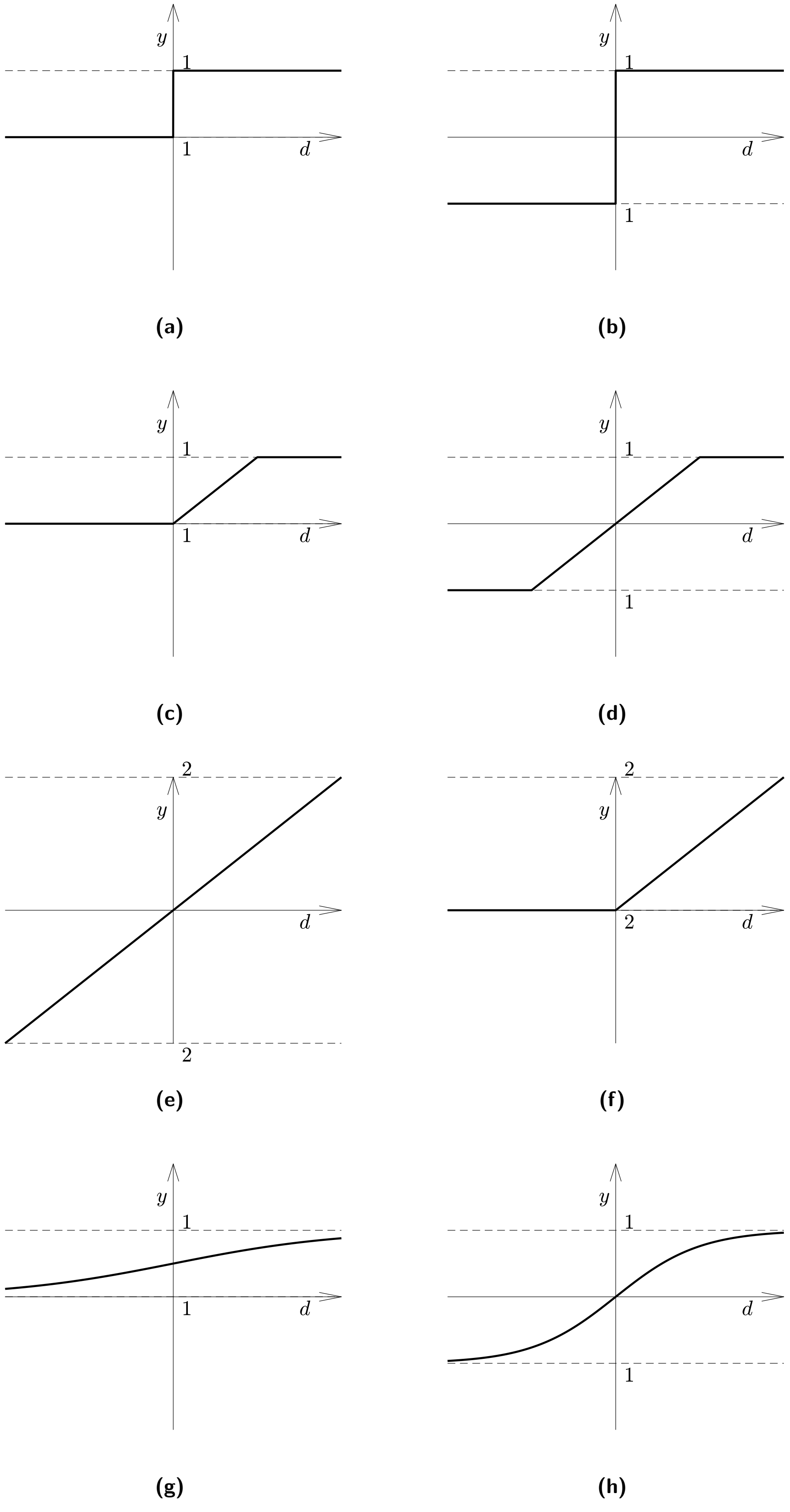


Figure A.3: Some of the possible activation functions for ANNs: asymmetrical (a) and symmetrical hard limit (b), asymmetrical (c) and symmetrical linear saturation (d), linear (e), rectified linear (f), standard logistic sigmoid (g) and hyperbolic tangent functions (h).

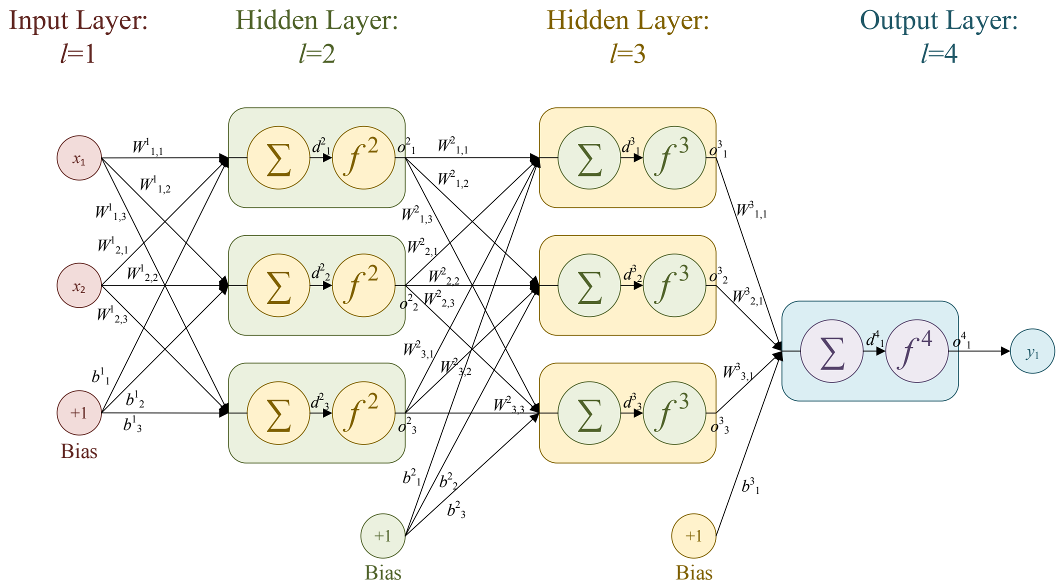


Figure A.4: Diagram of the ANN in Figure A.1 expressed with the mathematical model of Figure A.2.

samples, is employed *after* the network is trained to assess its performance on unseen points. The purpose for this check is to assess the importance of over-fitting, i.e. the mapping of not only the underlying relationship between input and output data, but also of the inevitable noise (Ng, 2016). If the performance of the ANN is excellent on the training set, but substantially worse for the validation set, over-fitting is an issue with the network and the network geometry should be modified, by reducing the number of neurons or by changing the activation functions or training method. Therefore, the validation set is employed to select the best performing approach. Finally, the test set, which entails 10%-25% of the original data samples, has a similar purpose to the validation set. However, it is employed to more specifically estimate the accuracy of the chosen approach after the method has been finalised. As a result, test data sets are often used to compare the performance of different machine learning algorithms (Ng, 2016).

In the next section, we describe the procedure for the computation of the predicted output given the input and weight matrices for each sample through a process known as forward propagation. In Section A.2.4, backpropagation is used to train the weights of the ANN.

A.2.3 Forward propagation

For each sample point k with input $\boldsymbol{x}_k$, it is possible to calculate the approximate output $\boldsymbol{y}_k$ by propagating the signal forward through the network. This process, known as forward propagation, assumes the weight matrices $\boldsymbol{W}$ and $\boldsymbol{b}$ to have already been determined. Their training is explained in the next section.

The equations in Section A.2.1 are modified to reflect the multi-node and multi-layer nature of the ANN. For each layer $l = 2, \ldots, L$ with N^l nodes, the vectors of activation and output signals are computed as

$$\boldsymbol{d}^l = \boldsymbol{W}^{l-1^T} \boldsymbol{o}^{l-1} + \boldsymbol{b}^{l-1}, \tag{A.13a}$$

$$\boldsymbol{o}^l = f^l \left(\boldsymbol{d}^l \right), \tag{A.13b}$$

respectively. The matrices and vectors are discussed in the previous section and are shown graphically in Figure A.4. The weight and bias matrices $\boldsymbol{W}^l$ and $\boldsymbol{b}^l$ have size $\left(N^{l-1}, N^l\right)$ and $\left(N^l, 1\right)$, respectively. The signal vectors $\boldsymbol{d}^l$ and $\boldsymbol{o}^l$ have size $\left(N^l, 1\right)$. The input to the second layer is in fact equal to the input to the whole ANN for each sample point, i.e. $\boldsymbol{o}^1 = \boldsymbol{x}_k$. Similarly, the output of the output layer matches the output of the ANN, i.e. $\boldsymbol{y}_k = \boldsymbol{o}^L$.

A.2.4 Training: backpropagation

During the training process, the weights of the ANN are updated so as to provide the mapping of highest accuracy between input and output data in a specified training set that comprises K points. As aforementioned, different strategies have been developed over the years for the training of ANNs. Here, we consider one of the performance learning schemes: *backpropagation* (Hagan *et al.*, 1996). This algorithm is a generalization of the least mean-squares training scheme developed by Widrow and Hoff (1960). Thus, backpropagation can be considered as an approximate steepest descent algorithm, which presents the mean square error as performance index, whose adoption represents the key insight of Widrow and Hoff (1960). For a given training set with input matrix $\boldsymbol{x}_{\text{tr}}$ and output matrix $\boldsymbol{y}_{\text{tr}}$ with size $\left(K, N^1\right)$ and $\left(K, N^L\right)$, respectively, the corresponding output predicted by the ANN is computed with forward propagation as $\boldsymbol{y}$ with size $\left(K, N^L\right)$. N^1 indicates the number of features of the input layer, while N^L those of the output layer. The performance index employed by backpropagation is thus expressed as

$$J = \frac{1}{2} ||\boldsymbol{y}_{\text{tr}} - \boldsymbol{y}||^2. \tag{A.14}$$

The constant term 1/2 is typically included to cancel out the term 2 deriving from the differentiation of the cost function. Often, a weight decay term is included to help reduce over-fitting (Hagan *et al.*, 1996), which occurs when the ANN maps not only the underlying function, but also noise effects. The modified performance index with the addition of the *regularization* term is thus expressed as

$$J = \frac{1}{2}||\boldsymbol{y}_{\mathrm{tr}}(k) - \boldsymbol{y}(k)||^2 + \frac{\lambda}{2}\sum_{k=1}^{K}\sum_{l=2}^{L}\sum_{i=1}^{N^{l-1}}\sum_{j=1}^{N^l}\left(W_{i,j}^l(k)\right)^2. \tag{A.15}$$

Backpropagation consists in two main steps. Firstly, a forward sweep is run through the network, calculating the activation and output signals of each layer, including the predicted output of the ANN (LeCun *et al.*, 2015). This can be done using either a sample point or a batch of points at a time. Then, the second step consists in propagating a sensitivity signal that depends on the prediction error backwards through the network; hence, the name backpropagation. The propagated sensitivity signal in each node is dependent on how much each node contributes to the overall error and the weight and bias matrices are updated accordingly.

Although different learning strategies are effective in training networks with a single hidden layer, most techniques are incapable of dealing with deep ANNs. Backpropagation is an effective scheme that deals with this issue. Since its development in the 1980s (LeCun *et al.*, 2015), it has been fundamental in the rapid rise of neural network technologies. In fact, backpropagation can be considered as a practical application of the chain rule of differentiation for the calculation of the gradient of an objective function (LeCun *et al.*, 2015).

Even within the backpropagation framework, different practical implementations have been proposed for the training of the ANN weights. Here, two famous schemes are analysed: gradient descent and the Levenberg-Marquart algorithms. The former is described first due to its simpler nature. It updates the neural weights by employing one training sample at a time. Conversely, the Levenberg-Marquart method is an example of an efficient batch scheme, which employs multiple samples at a time for the update of the ANN.

A.2.4.1 Gradient descent

The gradient descent algorithm is a simple iterative scheme that updates the weights of the network using one sample at a time. The update is based on the minimization of the cost function in (A.15) at every iteration a. Using a steepest descent algorithm (Hagan *et al.*, 1996), it is possible to express the update of the weights as

$$W_{i,j}^l(a+1) = W_{i,j}^l(a) - \alpha\frac{\partial J}{\partial W_{i,j}^l}, \tag{A.16a}$$

$$b_i^l(a+1) = b_i^l(a) - \alpha\frac{\partial J}{\partial b_i^l}, \tag{A.16b}$$

for every iteration a and sample k. $\alpha \in [0, 1]$ is known as the learning rate and it determines the quality of the learning.

In matrix form, for each sample k the gradient descent algorithm for backpropagation can be thus summarized as follows (Ng, 2016):

- Run a feed-forward sweep through the neural network, computing the activations for layers $l = 2, \dots, L$. Use (A.13a) and (A.13b).
- For the output layer L, obtain the error or sensitivity term $\boldsymbol{\delta}$, which is a measure of the influence of each neuron on any errors in the predicted output, as

$$\boldsymbol{\delta}^L = -\left(\boldsymbol{y}_{\mathrm{tr}}(k) - \boldsymbol{y}(k)\right) \odot \dot{f}^L\left(\boldsymbol{d}^l\right), \tag{A.17}$$

  for every sample k, where the dot indicates differentiation and $\odot$ the Hadamard, or element-wise, product.
- The error signal is then propagated backwards along the network for every layer $l = L-1, L-2, \dots, 2$:

$$\boldsymbol{\delta}^l = \left[\left(\boldsymbol{W}^l\right)^T \boldsymbol{\delta}^{l+1}\right] \odot \dot{f}^L\left(\boldsymbol{d}^l\right). \tag{A.18}$$

- The change in the partial derivatives of the weight and bias matrices is computed as:

$$\nabla_{\boldsymbol{W}^l} J = \boldsymbol{\delta}^{l+1}\left(\boldsymbol{o}^l\right)^T, \tag{A.19a}$$

$$\nabla_{\boldsymbol{b}^l} J = \boldsymbol{\delta}^{l+1}. \tag{A.19b}$$

- Add the regularization term.
- Update the weight and bias matrices with (A.16a) and (A.16b).

The procedure is repeated for a number of epochs a until convergence is achieved. This is typically determined by the change in the mean squared error on the training data set being less than a predefined value (Hagan *et al.*, 1996).

A.2.4.2 Levenberg-Marquardt algorithm

In mathematics, the Levenberg-Marquardt algorithm can be considered to be an evolution of the Newton-Gauss method for the minimization of functions that consist in the sum of quadratic terms, which is itself an evolution of Newton's method. Greater information on non-linear optimization schemes can be found in Scales (1985), as this topic goes beyond the scope of this thesis. The evolution of Newton's method into the Levenberg-Marquardt technique is also described in Chapter 12 of Hagan *et al.* (1996). The treatment of a quadratic cost function is fundamental in improving the algorithm performance, as it removes the need for the determination of the second derivative of the cost function (Hagan *et al.*, 1996).

In a minimization framework, the goal is to find the minimum point of the function $\boldsymbol{f}(\boldsymbol{x})$ quadratic in the vector of independent variables $\boldsymbol{x}$:

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{v}^T(\boldsymbol{x})\boldsymbol{v}(\boldsymbol{x}). \tag{A.20}$$

Using the Levenberg-Marquardt scheme, at every step a it is possible to update the vector of independent variables as follows (Hagan *et al.*, 1996):

$$\boldsymbol{x}_{a+1} = \boldsymbol{x}_a - \left[\boldsymbol{J}^T\left(\boldsymbol{x}_a\right)\boldsymbol{J}\left(\boldsymbol{x}_a\right) + \mu_a\boldsymbol{I}\right]^{-1}\boldsymbol{J}^T\left(\boldsymbol{x}_a\right)\boldsymbol{v}\left(x_a\right), \tag{A.21}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{J}\left(\boldsymbol{x}\right)$ is the Jacobian of the vector of independent variables. The term μ is fundamental in ensuring that the Hessian matrix of $\boldsymbol{x}$ is invertible. The Levenberg-Marquardt algorithm converges to the steepest gradient descent with small learning rate as μ is increased (Hagan *et al.*, 1996). Conversely, the algorithm becomes the Gauss-Newton scheme for $\mu = 0$. By starting with a small trial value for μ and then increasing it based on the convergence properties, the method is a good compromise between the speed of Newton's approach and the guarantee of convergence associated with steepest descent (Hagan *et al.*, 1996).

In the application to neural networks, the function to be minimized is the quadratic cost function in (A.15). In addition, the problem is extended to the treatment of multiple samples of the training set, as implied by the category batch learning under which the Levenberg-Marquardt scheme is classified. The application of the Levenberg-Marquardt to the training of neural networks is straightforward and similar to the approach described in the previous section for steepest gradient descent. However, the equations of the sensitivity of each layer need to be slightly modified. A detailed treatment of this method goes beyond the scope of this work, as the derivation is rather long. Therefore, the reader is referred to the article by Hagan and Menhaj (1994), which presents the first application of the Levenberg-Marquardt algorithm to the training of ANNs, and Chapter 12 of Hagan *et al.* (1996), which contains a clearer summary.

Bibliography

Abbeel, P. *Apprenticeship Learning and Reinforcement Learning with Application to Robotic Control.* PhD thesis, 2008.

Abraham, E. and Kerrigan, E. Optimal active control and optimization of a wave energy converter. {...} *Energy, IEEE Transactions on*, 4(2):1–8, 2013. URL `http://ieeexplore.ieee.org/xpls/abs{%}7B{_}{%}7Dall.jsp?arnumber=6353624`.

Amann, K. U., Magaña, M. E., Member, S., and Sawodny, O. Model Predictive Control of a Nonlinear 2-Body Point Absorber Wave Energy Converter With Estimated State Feedback. *Sustainable Energy, IEEE Transactions on*, 6(2):336–345, 2015.

Andersen, P., Pedersen, T. S., Nielsen, K. M., and Vidal, E. Model Predictive Control of a Wave Energy Converter. In *IFAC Workshop Series*, volume 19, pages 1540–1545, 2014. ISBN 9781479977864.

Arora, J. S. *Introduction to Optimum Design.* Academic Press, 3rd edition, 2012. ISBN 9780123813756. doi: 10.1016/B978-0-12-381375-6.00009-7.

Atiya, A. *Learning Algorithms for Neural Networks.* Phd thesis, California Insitute of Technology, 1991.

Babarit, A. and Clément, A. H. Optimal latching control of a wave energy device in regular and irregular waves. *Applied Ocean Research*, 28(2):77–91, 2006. ISSN 01411187. doi: 10.1016/j.apor.2006.05.002.

Babarit, A., Duclos, G., and Clément, A. H. Comparison of latching control strategies for a heaving wave energy device in random sea. *Applied Ocean Research*, 26(5): 227–238, 2004. ISSN 01411187. doi: 10.1016/j.apor.2005.05.003.

Babarit, A., Hals, J., Muliawan, M. J., Kurniawan, A., Moan, T., and Krokstad, J. Numerical benchmarking study of a selection of wave energy converters. *Renewable Energy*, 41:44–63, 2012. ISSN 09601481. doi: 10.1016/j.renene.2011.10.002.

Babarit, A. and Delhommeau, G. Theoretical and numerical aspects of the open source BEM solver NEMOH. *Proceedings of the 11th European Wave and Tidal Energy Conference.*, (September 2015):1–12, 2015. doi: hal-01198800.

Babarit, A., Guglielmi, M., and Clément, A. H. Declutching control of a wave energy converter. *Ocean Engineering*, 36(12-13):1015–1024, 2009. ISSN 00298018. doi: 10.1016/j.oceaneng.2009.05.006.

Bacelli, G. *Optimal control of wave energy converters.* Phd, NUI Maynooth, 2014.

Bellman, R. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

Bertsekas, D. P. and Tsitsiklis, J. *Neuro-Dynamic Programming.* Athena Scientific, hardcover edition, 1996. ISBN 1-886529-10-8.

Bhinder, M. a., Babarit, A., Gentaz, L., and Ferrant, P. Assessment of Viscous Damping via 3D-CFD Modelling of a Floating Wave Energy Device. *European Wave and Tidal Energy Conference, EWTEC*, pages 1–6, 2011.

Bjarte-Larsson, T. and Falnes, J. Laboratory experiment on heaving body with hydraulic power take-off and latching control. *Ocean Engineering*, 33(7):847–877, 2006. ISSN 00298018. doi: 10.1016/j.oceaneng.2005.07.007.

Bordons, C. and Camacho, E. F. *Model Predictive Control.* Springer-Verlag, 2nd edition, 2007. ISBN 978-1852336943.

Bradtke, S. J., Barto, A. G., and Kaelbling, P. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 57:33–57, 1996. ISSN 0885-6125. doi: 10.1007/BF00114723.

Brekken, T. K. A. On Model Predictive Control for a point absorber Wave Energy Converter. *Proceedings of the IEEE Trondheim PowerTech*, pages 1–8, 2011. doi: 10.1109/PTC.2011.6019367.

Budal, K. and Falnes, J. Optimum Operation of Wave Power Converter. *Marine Science Communications*, 3(2):133–150, 1977.

Busoniu, L., Babuska, R., Schutter, B. D., Ernst, D., Busoniu, L., Babuska, R., Schutter, B. D., and Ernst, D. *Reinforcement learning and dynamic programming using function approximators.* CRC Press, first edition, 2010. ISBN 978-1-4398-2108-4. doi: 10.1201/9781439821091.

Byrd, R. H., Gilbert, J. C., and Nocedal, J. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming, Series B*, 89(1):149–185, 2000. ISSN 00255610. doi: 10.1007/s101070000189.

Cantarellas, A. M., Remon Rodriguez, D., and Rodriguez, P. Adaptive Vector Control of Wave Energy Converters. *IEEE Transactions on Industry Applications*, 9994(c): 1–1, 2017. ISSN 0093-9994. doi: 10.1109/TIA.2017.2655478.

Castellini, L., Andrea, M. D., and Borgarelli, N. Analysis and Design of a Recipracating Linear Generator for a PTO. In *International Symposium on Power Electronics, Electrical Drives, Automation and Motion Analysis*, pages 1373–1379, 2014. ISBN 9781479947492.

Clément, A. H. and Babarit, A. Discrete control of resonant wave energy devices. *Philosophical Transactions of the Royal Society A*, 370:288–314, 2012. doi: 10.1098/rsta.2011.0132.

Combourieu, A., Philippe, M., Rongère, F., and Babarit, A. InWave : A New Flexible Design Tool Dedicated to Wave Energy Converters. *Volume 9B: Ocean Renewable Energy*, page V09BT09A050, 2014. doi: 10.1115/OMAE2014-24564. URL `http://proceedings.asmedigitalcollection.asme.org/proceeding.aspx?doi=10.1115/OMAE2014-24564`.

Cretel, J. a. M., Lightbody, G., Thomas, G. P., and Lewis, a. W. Maximisation of energy capture by a wave-energy point absorber using model predictive control. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 18(PART 1):3714–3721, 2011. ISSN 14746670. doi: 10.3182/20110828-6-IT-1002.03255.

Cruz, J. *OceanWave Energy.* Springer-Verlag, 2008.

Cummins, W. E. The impulse response function and ship motions. *Schiffstechnik*, 47 (9):101–109, 1962.

Danielsson, O. *Wave Energy Conversion: Linear Synchronous Permamnet Magnet Generator.* Phd, Uppsala University, 2006.

Davidson, J., Giorgi, S., and Ringwood, J. V. Identification ofWave Energy Device Models From NumericalWave Tank DataâĂŤPart 1: Numerical Wave Tank Identification Tests. *IEEE Transactions on Sustainable Energy*, 7(3):1012–1019, 2016. ISSN 1949-3029. doi: 10.1109/TSTE.2016.2515500.

Dennis, J. E. and Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, 1983.

Det Norske Veritas. Environmental conditions and environmental loads. *Recommended Practice*, (October):9–123, 2010.

Drew, B., Plummer, A., and Sahinkaya, M. N. A review of wave energy converter technology. *Journal of Power and Energy*, 223:887–902, 2009. ISSN 0957-6509. doi: 10.1243/09576509JPE782.

Elman, J. L. Finding Structure in Time. *Cognitive science*, 14(2):179–211, 1990. ISSN 03640213. doi: 10.1207/s15516709cog1402_1.

Eriksson, M., Waters, R., Svensson, O., Isberg, J., and Leijon, M. Wave power absorption: Experiments in open sea and simulation. *Journal of Applied Physics*, 102(8), 2007. ISSN 00218979. doi: 10.1063/1.2801002.

Eriksson, M. *Modelling and Experimental Verification of Direct Drive Wave Energy Conversion*. Phd, Uppsala University, 2007.

Eriksson, M., Isberg, J., and Leijon, M. Theory and experiment on an elastically moored cylindrical buoy. *IEEE Journal of Oceanic Engineering*, 31(4):959–963, 2006. ISSN 03649059. doi: 10.1109/JOE.2006.880387.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-Based Batch Mode Reinforcement Learning. *Journal of Machine Learning Research*, 6(1):503–556, 2005. ISSN 15324435.

Falcão, A. F. D. O. Phase control through load control of oscillating-body wave energy converters with hydraulic PTO system. *Ocean Engineering*, 35(3-4):358–366, 2008. ISSN 00298018. doi: 10.1016/j.oceaneng.2007.10.005.

Falcão, A. F. D. O. Wave energy utilization: A review of the technologies. *Renewable and Sustainable Energy Reviews*, 14(3):899–918, 2010. ISSN 13640321. doi: 10.1016/j.rser.2009.11.003.

Falcão, A. F. O. and Henriques, J. C. C. Oscillating-water-column wave energy converters and air turbines: A review. *Renewable Energy*, 85:1391–1424, 2016. ISSN 18790682. doi: 10.1016/j.renene.2015.07.086.

Falnes, J. Radiation impedance matrix and optimum power absorption for interacting oscillators in surface waves. *Applied Ocean Research*, 2(2):75–80, 1980. ISSN 01411187. doi: 10.1016/0141-1187(80)90032-2.

Falnes, J. On non-causal impulse response functions related to propagating water waves. *Applied Ocean Research*, 17(6):379–389, 1995. ISSN 01411187. doi: 10.1016/S0141-1187(96)00007-7.

Falnes, J. *Ocean waves and Oscillating systems*. Cambridge University Press, paperback edition, 2005. ISBN 0511030932. doi: 10.1016/S0029-8018(02)00070-7.

Falnes, J. A review of wave-energy extraction. *Marine Structures*, 20(4):185–201, 2007. ISSN 09518339. doi: 10.1016/j.marstruc.2007.09.001.

Farley, F., Rainey, R., and Chaplin, J. The peaks and troughs of wave energy: the dreams and the reality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1959):201–529, 2012.

Ferri, F., Ambühl, S., Fischer, B., and Kofoed, J. P. Balancing power output and structural fatigue ofwave energy converters by means of control strategies. *Energies*, 7(4):2246–2273, 2014. ISSN 19961073. doi: 10.3390/en7042246.

Forehand, D., Kiprakis, A. E., Nambiar, A., and Wallace, R. A Bi-directional Wave-to-Wire Model of an Array of Wave Energy Converters. *IEEE Transactions on Sustainable Energy*, 7(1):118–128, 2016. ISSN 19493029. doi: 10.1109/TSTE.2015.2476960.

Franklin, G. F., Powell, J. D., and Emami-Naeini, A. *Feedback Control of Dynamic Systems*. Pearson, 6th editio edition, 2008. ISBN 978-0135001509.

Fusco, F. and Ringwood, J. Short-Term Wave Forecasting for time-domain Control of Wave Energy Converters. *IEEE Transactions on Sustainable Energy*, 1(2):99–106, 2010a.

Fusco, F. and Ringwood, J. V. Short-term wave forecasting with ar models in real-time optimal control of wave energy converters. *IEEE International Symposium on Industrial Electronics*, pages 2475–2480, 2010b. ISSN 19493029. doi: 10.1109/ISIE.2010.5637714.

Fusco, F. and Ringwood, J. V. A simple and effective real-time controller for wave energy converters. *IEEE Transactions on Sustainable Energy*, 4(1):21–30, 2013. ISSN 19493029. doi: 10.1109/TSTE.2012.2196717.

Fusco, F. and Ringwood, J. V. Hierarchical robust control of oscillating wave energy converters with uncertain dynamics. *IEEE Transactions on Sustainable Energy*, 5 (3):958–966, 2014. ISSN 19493029. doi: 10.1109/TSTE.2014.2313479.

Geramifard, A., J.Walsh, T., Tellex, S., Chowdhary, G., Roy, N., and How, J. P. A Tutorial on Linear Function Approximators for Dynamic Programming and Reinforcement Learning. *Foundations and Trends{®} in Machine Learning*, 6(4): 375–451, 2013. ISSN 1935-8237. doi: 10.1561/2200000042.

Gieske, P. *Model predictive control of a wave energy converter: Archimedes Wave Swing*. M.sc., Delft University of Technology, 2007.

Giorgi, G., Pe, M., and Ringwood, J. V. Nonlinear Hydrodynamic Models for Heaving Buoy Wave Energy Converters. In *3rd Asian Wave and Tidal Energy Conference*, number October, Singapore, 2016a.

Giorgi, S., Davidson, J., and Ringwood, J. V. Identification of Wave Energy Device Models From Numerical Wave Tank DataâĂŤPart 2: Data-Based Model Determination. *IEEE Transactions on Sustainable Energy*, 7(3):1020–1027, 2016b. ISSN 1949-3029. doi: 10.1109/TSTE.2016.2515500.

Gunn, K. and Stock-Williams, C. Quantifying the Potential Global Market for Wave Power. *Proceedings of the 4th International Conference on Ocean Engineering (ICOE 2012)*, pages 1–7, 2012.

Hagan, M. T., Demuth, H. B., Beale, M. H., and De Jesús, O. *Neural Network Design*. PWS Publishing, second edition, 1996. ISBN 9780971732100.

Hagan, M. T. and Menhaj, M. B. Training Feedforward Networks with the Marquardt Algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993, 1994. ISSN 19410093. doi: 10.1109/72.329697.

Hals, J., Falnes, J., and Moan, T. Constrained Optimal Control of a Heaving Buoy Wave-Energy Converter. *Journal of Offshore Mechanics and Arctic Engineering*, 133 (1):11401, 2011. ISSN 08927219. doi: 10.1115/1.4001431.

Hansen, R. H., Kramer, M. M., and Vidal, E. Discrete displacement hydraulic power take-off system for the wavestar wave energy converter. *Energies*, 6(8):4001–4044, 2013. ISSN 19961073. doi: 10.3390/en6084001.

Harnois, V., Weller, S. D., Johanning, L., Thies, P. R., Le Boulluec, M., Le Roux, D., Soulè, V., and Ohana, J. Numerical model validation for mooring systems: Method and application for wave energy converters. *Renewable Energy*, 75:869–887, 2015. ISSN 18790682. doi: 10.1016/j.renene.2014.10.063.

Hebb, D. The Organization of Behavior. A neuropsychological theory. *The Organization of Behavior*, 911(1):335, 1949. ISSN 03619230. doi: 10.2307/1418888.

Henderson, R. Design, simulation, and testing of a novel hydraulic power take-off system for the Pelamis wave energy converter. *Renewable Energy*, 31(2):271–283, 2006. ISSN 09601481. doi: 10.1016/j.renene.2005.08.021.

Henriques, J. C. C., Gato, L. M. C., Falcão, A. F. O., Robles, E., and Faÿ, F. X. Latching control of a floating oscillating-water-column wave energy converter. *Renewable Energy*, 90:229–241, 2016. ISSN 18790682. doi: 10.1016/j.renene.2015.12.065.

Holthuijsen, L. H. *Waves in Oceanic and Coastal Waters*. Cambridge University Press, 2007. ISBN 978-0-521-86028-4.

Hoskin, R. E. and Nichols, N. K. Optimal strategies for phase control of wave energy devices. In McCormick, M. E. and Kim, Y. C., editors, *Utilization of ocean waves: wave to energy conversion*, pages 184–199, New York, NY, 1986. American Society of Civil Engineering.

Howard, R. A. *Dynamic Programming and Markov Processes*. Massachusetts Institute of Technology Press, 4 edition, 1960.

Jaakkola, T., Jordan, M. I., and Singh, S. P. On the Convergence of Stochastic Iterative Dynamic Programming Algorithms. *Neural Computation*, 6(6):1993, 1994.

Khan, S. G., Herrmann, G., Lewis, F. L., Pipe, T., and Melhuish, C. Reinforcement learning and optimal adaptive control: An overview and implementation examples. *Annual Reviews in Control*, 36(1):42–59, 2012. ISSN 13675788. doi: 10.1016/j.arcontrol.2012.03.004.

Koller, D. and Parr, R. Policy Interation for Factored MDPs. In *Uncertainty in Artificial Intelligence*, pages 326–334, 2000. ISBN 9780874216561. doi: 10.1007/s13398-014-0173-7.2.

Korde, U. A. Latching control of deep water wave energy devices using an active reference. *Ocean Engineering*, 29(11):1343–1355, 2002. ISSN 00298018. doi: 10.1016/S0029-8018(01)00093-2.

Korde, U. A. and Ringwood, J. V. *Hydrodynamic Control of Wave Energy Devices.* Cambridge University Press, Cambridge, 2016. ISBN 9781107079700.

Korde, U. A., Robinett, R. D., and Wilson, D. G. Wave-by-wave control in irregular waves for a wave energy converter with approximate parameters. *Journal of Ocean Engineering and Marine Energy*, 2(4):501–519, 2016. ISSN 2198-6444. doi: 10.1007/s40722-016-0068-0.

Kristiansen, E., Hjulstad, A., and Egeland, O. State-space representation of radiation forces in time-domain vessel models. *Modeling, Identification and Control*, 27(1): 23–41, 2006. ISSN 03327353. doi: 10.1016/j.oceaneng.2005.02.009.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003. ISSN 15324435. doi: 10.1162/jmlr.2003.4.6.1107.

Lawson, M., Yu, Y.-H., Nelessen, A., Ruehl, K., and Michelen, C. Implementing nonlinear buoyancy and excitation forces in the WEC-Sim wave energy converter modeling tool. In *33rd International Conference on Ocean, Offshore and Arctic Engineering*, pages 1–6, San Francisco, 2014. ASME.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 0028-0836. doi: 10.1038/nature14539.

Lejerskog, E., Boström, C., Hai, L., Waters, R., and Leijon, M. Experimental results on power absorption from a wave energy converter at the Lysekil wave energy research site. *Renewable Energy*, 77:9–14, 2015. ISSN 18790682. doi: 10.1016/j.renene.2014.11.050.

Leontaritis, I. J. and Billings, S. A. Input-output parametric models for non-linear systems Part I: deterministic non-linear systems. *International Journal of Control*, 41(2):303–328, 1985a.

Leontaritis, I. J. and Billings, S. A. Input-output parametric models for non-linear systems Part II: stochastic non-linear systems. *International Journal of Control*, 41 (2):329–344, 1985b.

Levy, E. Complex-curve fitting. *IRE Transactions on Automatic Control*, AC-4(1): 37–43, 1959. ISSN 0096-199X. doi: 10.1109/TAC.1959.6429401.

Lewis, F., Vrabie, D., and Vamvoudakis, K. Reinforcement Learning and Feedback Control: Using Natural Decision Methods to Design Optimal Adaptive Controllers. *IEEE Control Systems*, 32(6):76–105, 2012. ISSN 1066-033X. doi: 10.1109/MCS. 2012.2214134.

Li, G. and Belmont, M. R. Model predictive control of sea wave energy converters - Part I: A convex approach for the case of a single device. *Renewable Energy*, 69: 453–463, 2014a. ISSN 09601481. doi: 10.1016/j.renene.2014.03.070.

Li, G. and Belmont, M. R. Model predictive control of sea wave energy converters - Part II: The case of an array of devices. *Renewable Energy*, 68:540–549, 2014b. ISSN 09601481. doi: 10.1016/j.renene.2014.02.028.

Li, G., Weiss, G., Mueller, M., Townley, S., and Belmont, M. R. Wave energy converter control by wave prediction and dynamic programming. *Renewable Energy*, 48:392–403, 2012.

Li, Y. and Yu, Y.-H. A synthesis of numerical methods for modeling wave energy converter-point absorbers. *Renewable and Sustainable Energy Reviews*, 16(6):4352–4364, 2012. ISSN 13640321. doi: 10.1016/j.rser.2011.11.008.

Littman, M. L. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521(7553):445–451, 2015. ISSN 1476-4687. doi: 10.1038/nature14540. URL `http://www.nature.com/doifinder/10.1038/nature14540{%}7B{%}25{%}7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26017443`.

Lok, K. S., Stallard, T. J., Stansby, P. K., and Jenkins, N. Optimisation of a clutch-rectified power take off system for a heaving wave energy device in irregular waves with experimental comparison. *International Journal of Marine Energy*, 8:1–16, 2014. ISSN 22141669. doi: 10.1016/j.ijome.2014.09.001.

Lucas, J., Livingstone, M., Vuorinen, M., and Cruz, J. Development of a wave energy converter (WEC) design tool - application to the WaveRoller WEC including validation of numerical estimates. *Proceedings of 4th International Conference on Ocean Energy*, 2012.

Mandic, D. P. and Chambers, J. A. *Recurrent Neural networks for prediction-learning algorithms, architectures and stability*. John Wiley & Sons, Chichester, 2001. ISBN 978-0-471-49517-8.

Maslin, M. *Climate Change: A Very Short Introduction*. Oxford University Press, Oxford, third edition, 2014.

McCulloch, W. S. and Pitts, W. A Logical Calculus of the Idea Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943. ISSN 0007-4985. doi: 10.1007/BF02478259.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. a., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015. ISSN 0028-0836. doi: 10.1038/nature14236. URL `http://dx.doi.org/10.1038/nature14236`.

Moody, J. E. and Darken, C. Fast Learning in Networks of Locally-Tuned Processing Units. *Neural Computation*, 1(2):281–294, 1989.

Morison, J. R., Johnson, J. W., and Schaaf, S. The Force Exerted by Surface Waves on Piles. *Journal of Petroleum Technology*, 2(5):149–154, 1950. ISSN 0149-2136. doi: 10.2118/950149-G.

Munos, R. Error Bounds for Approximate Policy Iteration. *Proceedings, Twentieth International Conference on Machine Learning*, 2:560–567, 2003. URL `http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-1942516880{%}7B{&}{%}7DpartnerID=40{%}7B{&}{%}7Drel=R8.2.0`.

Nambiar, A. J., Forehand, D. I. M., Kramer, M. M., Hansen, R. H., and Ingram, D. M. Effects of hydrodynamic interactions and control within a point absorber array on electrical output. *International Journal of Marine Energy*, 9:20–40, 2015. ISSN 2214-1669. doi: 10.1016/j.ijome.2014.11.002.

Neary, V. S., Previsic, M., Jepsen, R. a., Lawson, M. J., Yu, Y.-H., Copping, A. E., Fontaine, A. a., and Hallett, K. C. Methodology for Design and Economic Analysis of Marine Energy Conversion (MEC) Technologies. Technical Report March, Sandia National Laboratories, 2014.

Newman, J. N. The Exciting Forces on Fixed Bodies in Waves. *Journal of Ship Research, Society of Naval Architects and Marine Engineers*, 6:1–10, 1962.

Newman, J. N. *Marine Hydrodynamics*. MIT Press, 1977. ISBN 9780262140263.

Ng, A. Machine Learning Course Materials, 2016. URL `http://cs229.stanford.edu/materials.html`.

Nørgaard, M., Ravn, O., Poulsen, N. K., and Hansen, L. K. *Neural networks for modelling and control of dynamic systems: A practitioner's handbook*. Springer-Verlag, London, 2003.

NREL. WEC-Sim (Wave Energy Converter SIMulator), 2015. URL `http://wec-sim.github.io/WEC-Sim/index.html`.

Oetinger, D., Magaña, M. E., Member, S., and Sawodny, O. Decentralized Model Predictive Control for Wave Energy Converter Arrays. *Sustainable Energy, IEEE Transactions on*, 5(4):1099–1107, 2014.

Oetinger, D., Magaña, M. E., and Sawodny, O. Centralised model predictive controller design for wave energy converter arrays. *IET Renewable Power Generation*, 9(2): 142–153, 2015. ISSN 1752-1416. doi: 10.1049/iet-rpg.2013.0300.

Ogilvie, T. Recent progress toward the understanding and prediction of ship motions. Fifth Symposium on Naval Hydrodynamics, Bergen, Norway, 1964.

Peñalba Retes, M. and Ringwood, J. V. A review of wave-to-wire models for wave energy converters. *Energies*, 9(7), 2016. ISSN 19961073. doi: 10.3390/en9070506.

Peñalba Retes, M., Giorgi, G., and Ringwood, J. V. A Review of non-linear approaches for wave energy converter modelling. *11th European Wave and Tidal Energy Conference*, (1):1–10, 2015.

Peñalba, M., Giorgi, G., and Ringwood, J. V. Mathematical modelling of wave energy converters: A Review of nonlinear approaches. *Renewable and Sustainable Energy Reviews*, 78(August 2015):1188–1207, 2017. ISSN 18790690. doi: 10.1016/j.rser.2016.11.137.

Pérez, T. and Fossen, T. I. Time-vs. frequency-domain Identification of parametric radiation force models for marine structures at zero speed. *Modeling, Identification and Control*, 29(1):1–19, 2008. ISSN 03327353. doi: 10.4173/mic.2008.1.1.

Pizer, D. and Henderson, R. Pelamis Wave Energy Converter P2: PEL Simulation Overview. Technical report, Pelamis Wave Power Ltd., 2010.

Pontryagin, L. *Mathematical Theory of Optimal Processes*. CRC Press, 1987. ISBN 9782881240775.

Previsic, M., Shoele, K., and Epler, J. Validation of Theoretical Performance Results using Wave Tank Testing of Heaving Point Absorber Wave Energy Conversion Device working against a Subsea Reaction Plate. *2nd Marine Energy Technology Symposium*, pages 1–8, 2014.

Price, A. A. E. New Perspectives on Wave Energy Converter Control. (March), 2009.

Richter, M., Magana, M. E., Sawodny, O., and Brekken, T. K. a. Nonlinear Model Predictive Control of a Point Absorber Wave Energy Converter. *Sustainable Energy, IEEE Transactions on*, 4(1):118–126, 2013. doi: 10.1109/TSTE.2012.2202929.

Richter, M., Sawodny, O., Magaña, M. E., and Brekken, T. K. a. Power optimisation of a point absorber wave energy converter by means of linear model predictive control. *IET Renewable Power Generation*, 8(2):203–215, 2014. ISSN 1752-1416. doi: 10.1049/iet-rpg.2012.0214.

Riedmiller, M. Neural fitted Q iteration - First experiences with a data efficient neural Reinforcement Learning method. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3720 LNAI:317–328, 2005. ISSN 03029743. doi: 10.1007/11564096_32.

Riedmiller, M. 10 Steps and some tricks to set up neural reinforcement controllers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7700 LECTU:735–757, 2012. ISSN 03029743. doi: 10.1007/978-3-642-35289-8-39.

Riedmiller, M. and Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *IEEE International Conference on Neural Networks - Conference Proceedings*, 1993-Janua:586–591, 1993. ISSN 10987576. doi: 10.1109/ICNN.1993.298623.

Ringwood, J. V., Bacelli, G., and Fusco, F. Energy-Maximizing Control of Wave-Energy Converters: The Development of Control System Technology to Optimize Their Operation. *IEEE Control Systems Magazine*, 34(5):30–55, 2014.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519.

Ruehl, K., Michelen, C., Kanner, S., Lawson, M., and Yu, Y.-H. Preliminary Verification and Validation of WEC-SIM, an open-source wave energy converter design tool. In *33rd International Conference on Ocean, Offshore and Arctic Engineering*, pages 1–7, San Francisco, 2014. ASME.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representatons by back-propagating errors. *Nature*, 323(9):533–536, 1986. ISSN 0028-0836. doi: 10.1038/323533a0.

Rummery, G. A. and Niranjan, M. On-Line Q-Learning Using Connectionist Systems. Technical report, University of Cambridge, 1994.

Salter, S. H. Wave power. *Nature*, 249:720–724, 1974. ISSN 09637346. doi: 10.1049/esej:20000303.

Salter, S. H., Taylor, J. R. M., and Caldwell, N. J. Power conversion mechanisms for wave energy. *Proceedings of the I MECH E Part M*, 216(1):1–27, 2002. ISSN 14750902. doi: 10.1243/147509002320382112.

Santhosh, N., Baskaran, V., and Amarkarthik, A. A review on front end conversion in ocean wave energy converters. *Frontiers in Energy*, 9(3):297–310, 2015. ISSN 20951698. doi: 10.1007/s11708-015-0370-x.

Scales, L. E. *Introduction to Non-Linear Optimization*. Springer-Verlag, New-York, 1985.

Schmidhuber, J. Deep Learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. ISSN 18792782. doi: 10.1016/j.neunet.2014.09.003.

Schweitzer, P. J. and Seidmann, A. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110(2):568–582, 1985. ISSN 10960813. doi: 10.1016/0022-247X(85)90317-8.

Scruggs, J. T. and Nie, R. Disturbance-adaptive stochastic optimal control of energy harvesters, with application to ocean wave energy conversion. *Annual Reviews in Control*, 40:102–115, 2015. ISSN 13675788. doi: 10.1016/j.arcontrol.2015.09.017.

Shoori J., H. E., Ling, B., and Batten, B. A. Use of artificial neural networks for real-time prediction of heave displacement in ocean buoys. *3rd International Conference on Renewable Energy Research and Applications, ICRERA 2014*, pages 907–912, 2015. doi: 10.1109/ICRERA.2014.7016517.

Singh, S., Jaakkola, T., Littman, M. L., Szepes, C., and Hu, A. S. Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms. *Machine Learning*, 39(1998):287–308, 2000.

Sirnivas, S., Yu, Y.-H., Hall, M., and Bosma, B. Coupled Mooring Analyses for the WEC-Sim Wave Energy Converter Design Tool. *35th International Conference on Ocean, Offshore and Arctic Engineering*, (July), 2016. doi: 10.1115/OMAE2016-54789.

Stalberg, M., Waters, R., Danielsson, O., and Leijon, M. Influence of Generator Damping on Peak Power and Variance of Power for a Direct Drive Wave Energy Converter. *Journal of Offshore Mechanics and Arctic Engineering*, 130(3):1–4, 2008. ISSN 08927219. doi: 10.1115/1.2905032.

Stansell, P. and Pizer, D. J. Maximum wave-power absorption by attenuating line absorbers under volume constraints. *Applied Ocean Research*, 40:83–93, 2013. ISSN 01411187. doi: 10.1016/j.apor.2012.11.005.

Süli, E. and Mayers, D. *An Introduction to Numerical Analysis*. Cambridge University Press, Cambridge, paperback edition, 2003.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning*. MIT Press, hardcover edition, 1998.

Taghipour, R., Perez, T., and Moan, T. Hybrid frequency-time domain models for dynamic response analysis of marine structures. *Ocean Engineering*, 35(7):685–705, 2008. ISSN 00298018. doi: 10.1016/j.oceaneng.2007.11.002.

Thomas, P. and Evans, V. Arrays of three-dimensional wave-energy absorbers. *Journal of Fluid Mechanics*, 108(1981):67–88, 1981.

Titah-Benbouzid, H. and Benbouzid, M. Ocean wave energy extraction: Up-to-date technologies review and evaluation. *Proceedings - 2014 International Power Electronics and Application Conference and Exposition, IEEE PEAC 2014*, 2014. doi: 10.1109/PEAC.2014.7037878.

Tona, P., Nguyen, H.-n., Sabiron, G., and Creff, Y. An Efficiency-Aware Model Predictive Control Strategy for a Heaving Buoy Wave Energy Converter. *Proceedings of the 11th European Wave and Tidal Energy Conference*, pages 1–10, 2015.

Ugray, Z., Lasdon, L., Plummer, J., and Glover, F. Scatter Search and Local NLP Solvers : A Multistart Framework for Global Optimization. *Information Systems*, 19 (May):328–340, 2006. ISSN 1091-9856. doi: 10.1287/ijoc.1060.0175.

Valério, D., Beirão, P., and Sá da Costa, J. Optimisation of wave energy extraction with the Archimedes Wave Swing. *Ocean Engineering*, 34(17-18):2330–2344, 2007. ISSN 00298018. doi: 10.1016/j.oceaneng.2007.05.009.

Valério, D., Mendes, M. J. G. C., Beirão, P., and Sá da Costa, J. Identification and control of the AWS using neural network models. *Applied Ocean Research*, 30(3): 178–188, 2008. ISSN 01411187. doi: 10.1016/j.apor.2008.11.002.

Waltz, R. A., Morales, J. L., Nocedal, J., and Orban, D. An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical Programming*, 107(3):391–408, 2006. ISSN 00255610. doi: 10.1007/s10107-004-0560-5.

WAMIT. *User Manual: Version 7.0.* 2013. ISBN 9788578110796. doi: 10.1017/CBO9781107415324.004.

Waters, R. *Energy from Ocean Waves*. Phd, Uppsala University, 2008.

Waters, R., Engström, J., Isberg, J., and Leijon, M. Wave climate off the Swedish west coast. *Renewable Energy*, 34(6):1600–1606, 2009. ISSN 09601481. doi: 10.1016/j.renene.2008.11.016.

Watkins, C. J. *Models of Delayed Reinforcement Learning*. Ph.d., Cambridge University, 1989.

Watkins, C. J. C. H. and Dayan, P. Technical Note: Q-Learning. *Machine Learning*, 8 (3):279–292, 1992. ISSN 15730565. doi: 10.1023/A:1022676722315.

Wave Energy Scotland. Control Requirements for Wave Energy Converters Landscaping Study: Final Report. Technical report, Wave Energy Scotland, 2016.

Wave Energy Scotland. Control Systems Competition Specification and Guidance Document. Technical Report April, Wave Energy Scotland, Inverness, 2017.

Wei, C., Zhang, Z., Qiao, W., and Qu, L. Reinforcement learning-based intelligent maximum power point tracking control for wind energy conversion systems. *IEEE Transactions on Industrial Electronics*, 62(10):6360–6370, 2015. ISSN 0278-0046. doi: 10.1109/TIE.2015.2420792.

Wei, C., Zhang, Z., Qiao, W., and Qu, L. An Adaptive Network-Based Reinforcement Learning Method for MPPT Control of PMSG Wind Energy Conversion Systems. *IEEE Transactions on Power Electronics*, 8993(c):1, 2016. ISSN 0885-8993. doi: 10.1109/TPEL.2016.2514370.

Widrow, B. and Hoff, M. E. Adaptive Switching Circuits. Technical report, Stanford Electronics Laboratory, 1960.

Yu, Y., Lawson, M., Li, Y., Previsic, M., Epler, J., and Lou, J. Experimental Wave Tank Test for Reference Model 3 Floating- Point Absorber Wave Energy Converter Project. Technical Report January, National Renewable Energy Laboratory, 2015.

Zhu, X. and Lee, C.-H. Removing the Irregular Frequencies in Wave-Body Interactions. In *9th International Workshop for Water Waves and Floating Bodies*, pages 245–249, 1994. doi: 10.1017/S0022112089002636.

Zou, S., Abdelkhalik, O., Robinett, R., Bacelli, G., and Wilson, D. Optimal control of wave energy converters. *Renewable Energy*, 103:217–225, 2017. ISSN 18790682. doi: 10.1016/j.renene.2016.11.036.