

Deep Reinforcement Learning Based Controller for Active Heave Compensation

Shrenik Zinage* Abhilash Somayajula*

* *Department of Ocean Engineering, Indian Institute of Technology
Madras (IITM), Chennai, India, 600036 (e-mail:
shrenikvz@gmail.com, abhilash@iitm.ac.in)*

Abstract: Heave compensation is an essential part in various offshore operations. It is used in various applications, which include on-loading or off-loading systems, offshore drilling, landing helicopter on oscillating structures, and deploying and retrieving manned submersibles. In this paper, a reinforcement learning (RL) based controller is proposed for active heave compensation using a deep deterministic policy gradient (DDPG) algorithm. A DDPG algorithm which is a model-free, online reinforcement learning method, is adopted to capture the experience of the agent during the training trials. The simulation results demonstrate up to 10 % better heave compensation performance of RL controller as compared to a tuned Proportional-Derivative Control. The performance of the proposed method is compared with respect to heave compensation, offset tracking, disturbance rejection, and noise attenuation.

Keywords: deep reinforcement learning, learning-based control, winch control, artificial intelligence, active heave compensation

1. INTRODUCTION

With the increase in the number of ocean explorations and a huge demand for various marine resources, heave compensation has become a vital part of various maritime operations. In heave compensation, the primary objective is to decouple the motion of a payload connected to the ship from the ship's vertical (heave) motion. Heave compensation methods can be broadly classified into two main categories: passive heave compensation (PHC) and active heave compensation (AHC). The PHC is an open-loop system that is designed to partially decouple the payload from the vessel. The compensation performance for PHC is generally observed to be less than 80 % (Hatleskog and Dunnigan (2007)). AHC relies on closed-loop control system architecture and provides the payload displacement as a continuous feedback to the controller so that an improved compensation performance is achieved.

Over the years, various classical control techniques have been analysed and compared to keep a suspended payload regulated when the vessel is undergoing dynamic vertical motion in the ocean (Zinage and Somayajula (2020); Li et al. (2019); Woodacre et al. (2018); Do and Pan (2008)). However, not much research has looked at using reinforcement learning (RL) control techniques, which is one of the methods gaining popularity in marine applications (Woo et al. (2019); Martinsen and Lekkas (2018); Zhao and Roh (2019)).

Lately, model-free deep reinforcement learning has made significant progress in solving a variety of complex tasks. The first successful application of this technique to learn the control policy was through a deep Q-network (DQN) for playing Atari games (Mnih et al. (2013); Silver et al. (2017)), which integrates the Q learning and deep neu-

ral network. However, DQN can only be used to solve problems that have discrete action space. Since many control tasks in the real world have continuous action space, several advanced reinforcement learning algorithms aiming to solve continuous control problems have also been developed (Lillicrap et al. (2015); Mnih et al. (2016); Levine et al. (2016)).

In this paper, a deep deterministic policy gradient (DDPG) (Lillicrap et al. (2015)) algorithm that is based on an actor-critic framework is used. This method has an advantage over the DQN approach that it can deal with a continuous action space. Apart from that, this algorithm inherits conventional approaches of RL such as actor-critic (Sutton et al. (1999)), and policy gradient (Konda and Tsitsiklis (2000)).

Main Contributions: This paper introduces a RL based control methodology using the DDPG algorithm for active heave compensation. The efficacy and the potential of using this technique for real-life applications are assessed in a simulated environment. To the best of our knowledge, this is the first time that RL is being investigated as a tool for active heave compensation.

Structure of the paper: The organisation of the paper is as follows: Section 2 discusses the modeling of the ship's motion. The dynamic model of the winch is described in Section 3. Section 4 describes the theoretical background and problem formulation of RL based controller. The simulation results are presented in Section 5 and the conclusions are presented in Section 6.

2. MODELING OF SHIP MOTION

In this paper, the KRISO container ship (KCS) is chosen for calculating the dynamic motion in the ocean. Ta-

Table 1. KCS Particulars

Particulars	Value
Length between perpendiculars L_{pp}	230 m
Length waterline L_{WL}	232.5 m
Breadth B	32 m
Depth D	19 m
Draft T	10.8 m
Displacement	52030 m ³
Block Coefficient C_B	0.65
LCB from midship (fwd +)	-3.404 m
LCB from AP (fwd +)	111.596 m
VCG from WL	3.551 m
VCG from keel	14.351 m
GM	0.6 m
Design Forward Speed U	24 knots
Analysis Speed (In this study)	0 knots
Froude Number F_n	0.26
Roll Radius of gyration about CG k_{xx}	12.88 m
Pitch/Yaw radius of gyration about CG k_{yy}/k_{zz}	57.5 m

ble 1 shows the particulars of an KCS ship. The Pierson Moskowitz (PM) wave elevation spectrum corresponding to a significant wave height H_s and peak period T_p for a range of frequencies is defined by

$$S(\omega) = \frac{0.31}{2\pi} T_p H_s^2 \left(\frac{\omega T_p}{2\pi} \right)^{-5} \exp \left(-\frac{5}{4} \left(\frac{\omega T_p}{2\pi} \right)^4 \right) \quad (1)$$

An irregular wave elevation time history is generated from the above spectrum. The wave elevation is expressed as a sum of N sinusoidal components as shown below

$$\eta(t) = \sum_{i=1}^N A_n \cos(\omega_n t + \phi_n) \quad (2)$$

In this study, N is taken as 100001 to simulate a 10000 s time history with a sampling time of 0.1 s. For a simulation of duration T seconds, the frequency increment is given by $\Delta\omega = 2\pi/T$. At each discrete frequency $\omega_n = n\Delta\omega$, the amplitude of the n^{th} wave component is given by

$$A_n = \sqrt{2S(\omega_n)\Delta\omega} \quad (3)$$

The phase ϕ_n of each wave component is sampled from a uniform distribution between $-\pi$ and π (Somayajula (2017)). In order to avoid repeating of the generated signal the frequencies are randomised as shown below

$$(\omega_n)_{new} = (\omega_n)_{old} + \Delta\omega X \quad (4)$$

where X is a random variable following an uniform distribution between -0.5 and 0.5 .

The response amplitude operator (RAO) of the KCS container ship is obtained for the heave, roll and pitch modes using MDLHydroD developed by Guha (2016) which is a frequency domain 3D panel method based tool for analysis of wave structure interaction.

Once the RAO is obtained, the response spectrum of the roll, pitch and heave is calculated as shown below

$$S_{\text{response}}(\omega) = |H(\omega)|^2 S(\omega) \quad (5)$$

where S_{response} is the response spectrum, $H(\omega)$ is the RAO and $S(\omega)$ is the wave spectrum. Now the input wave elevation time history is decomposed into its frequency components by taking a fast Fourier transform (FFT). The amplitude of the response at a frequency $\omega_k = (k-1)\Delta\omega$ is then obtained by taking the product of RAO at that frequency and the FFT of input wave elevation time history at the same frequency. Finally, inverse fast Fourier

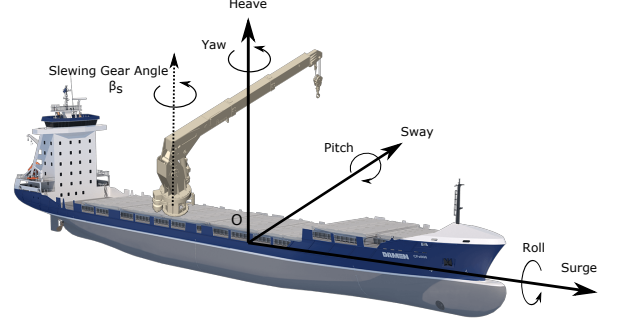


Fig. 1. Ship with Installed Crane

transform (IFFT) is used to convert these three degrees of freedom back into time domain.

In this study, the origin of the ship fixed coordinate system is assumed to be at the intersection of the waterline, centerline, and midship. Assuming that the crane is placed at $(x_{\text{crane}}, y_{\text{crane}})$ in the vessel's body fixed coordinate system with a slewing gear angle β_s and the wave incident angle of β , the net heave response time history of the winch placed on board the KCS ship is calculated in terms of the combined roll, heave, and pitch motion caused due to the wave excitation. Fig. 1 shows a schematic diagram of the ship with a crane installed on it. Assuming small amplitude motions consistent with linear hydrodynamic theory, the net heave motion time history is given by

$$z_{\text{winch}} = \eta_3(\beta) + (y_{\text{crane}} + l_{\text{crane}} \sin(\beta_s)) \eta_4(\beta) - (x_{\text{crane}} + l_{\text{crane}} \cos(\beta_s)) \eta_5(\beta) \quad (6)$$

where $\eta_3(\beta)$, $\eta_4(\beta)$, and $\eta_5(\beta)$ are the heave, roll and pitch time histories respectively, which depend on the incident wave angle β . In this study, the coordinate of the crane with respect to vessel's body frame is assumed to be at $(-1.5 \text{ m}, 2 \text{ m})$ with a slewing gear angle of 30 degrees and the horizontal extent of the crane (l_{crane}) is assumed to be

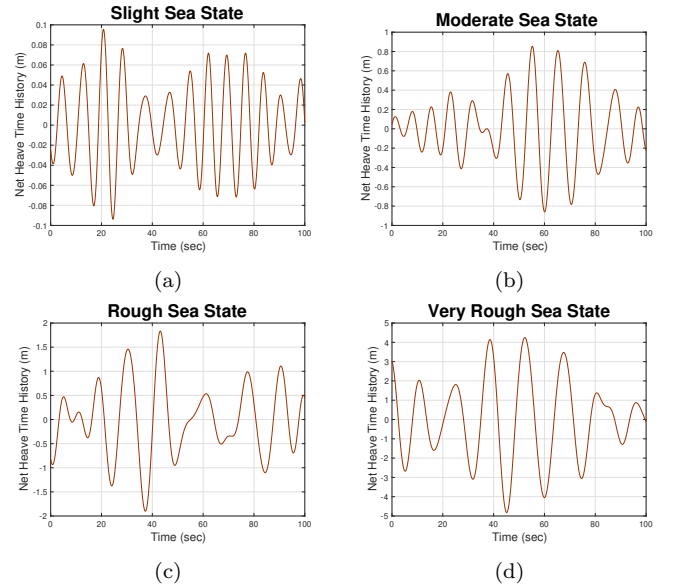


Fig. 2. Net heave time history generated from a PM spectra having (a) $H_s = 1.5 \text{ m}$, $T_p = 6 \text{ s}$ (b) $H_s = 4 \text{ m}$, $T_p = 9 \text{ s}$ (c) $H_s = 6 \text{ m}$, $T_p = 12 \text{ s}$ (d) $H_s = 8.5 \text{ m}$, $T_p = 14 \text{ s}$

3 m. The plot of a 100 second snip of the net heave motion time history in 4 different sea states when the waves are incident at an angle of 135 degrees are shown in Fig. 2.

3. STATE SPACE MODEL

In this section, the state space model for hydraulic drive of the winch is described (Richter et al. (2017)). A schematic representation of the hydraulic drive is shown in Fig. 3.

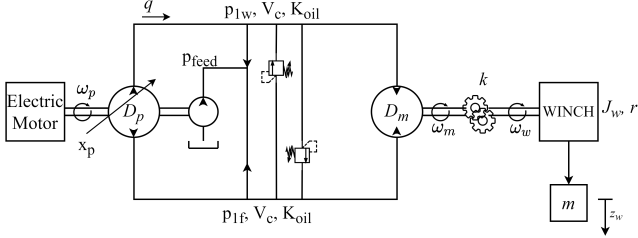


Fig. 3. Schematic diagram of hydraulic driven winch

The state space model of the winch is given by

$$\begin{aligned} \dot{x} &= Ax + Bu_p + [0 \ 0 \ d \ 0]^T \\ y &= z_w = Cx, \quad x(0) = x_0 \end{aligned} \quad (7)$$

with the state

$$x = [x_p \ \Delta p \ \dot{z}_w \ z_w]^T \quad (8)$$

where x_p , Δp , $\dot{z}_w$, z_w are normalized swash angle, change in pressure, velocity of the reeled rope, and length of the reeled rope respectively. The input to this plant is u_p that translates to the normalized swash angle through a first order system as shown in (9).

$$\dot{x}_p = \frac{-1}{T_w}(x_p - u_p) \quad (9)$$

A saturation of normalized swash angle is assumed beyond $x_p = \pm 1$. The system matrix is given by

$$A = \begin{bmatrix} -\frac{1}{T_w} & 0 & 0 & 0 \\ -\frac{2K_{oil}D_p\omega_p}{V_c} & -\frac{2K_{oil}k_{leak}}{V_c} & \frac{2K_{oil}D_mk}{rV_c} & 0 \\ 0 & -\frac{rD_mk\eta_m}{J_w + mr^2} & -\frac{b}{J_w + mr^2} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (10)$$

where $k_{leak} = k_{1,p} + k_{1,m}$, the input matrix by

$$B = \begin{bmatrix} \frac{1}{T_w} & 0 & 0 & 0 \end{bmatrix}^T \quad (11)$$

the output matrix by

$$C = [0 \ 0 \ 0 \ 1] \quad (12)$$

and disturbance by

$$d = \frac{r^2 mg}{J_w + mr^2} + \tilde{d} \quad (13)$$

where $\tilde{d}$ is the disturbance caused due to unmodelled dynamics, nonlinear friction, vibrations and parameter uncertainty. The parameters and their definitions for the hydraulic drive of the winch are shown in Table 2. Since the limits in pressure relief valve are not usually reached

in practise, the oil supply does not significantly affect the governing dynamics of the winch and hence the leakage of oil is not modeled in this paper.

A more comprehensive explanation of this model can be found in Zinage and Somayajula (2020). It is assumed that the cable does not lose tension throughout the operation and the net heave time history of the vessel is already known from predictions prior to the implementation of RL based control.

Table 2. Data Values

Parameter Name	Parameters	Value
Acceleration due to gravity	g	9.8 m/s^2
Bulk modulus of hydraulic fluid	K_{oil}	$1.8 \times 10^9 \text{ N/m}^2$
Volume of hydraulic lines	V_c	$2 \times 10^{-3} \text{ m}^3$
Maximum pump displacement	D_p	$40 \times 10^{-6} \text{ m}^3$
Fixed displacement of motor	D_m	$4 \times 10^{-6} \text{ m}^3$
Rotation rate of pump	ω_p	45 Hz
Leakage constant of pump	$k_{1,p}$	0
Leakage constant of motor	$k_{1,m}$	0
Time constant	T_w	1 s
Gear transmission ratio	k	200
Radius of winch	r	0.5 m
Efficiency of motor	η_m	0.65
Inertia of the winch	J_w	150 kgm^2
Viscous friction of the winch	b	$1 \times 10^4 \text{ kgm}^2/\text{s}$
Mass of the payload	m	1000 kg

4. REINFORCEMENT LEARNING BASED CONTROLLER

In RL, the agent interacts with the environment by taking actions and observing the state and the reward without any knowledge of the dynamics of the environment. The goal in RL algorithm is to find the optimal policy π^* that selects the optimal control actions u_t^* as shown below

$$u_t^* = \pi^*(s_t) = \arg \max_{\pi} Q^{\pi}(s_t, u_t) \quad (14)$$

that maximises the Q value, which is the expected value of the total discounted reward when an action u_t is taken in state s_t . Mathematically, it can be defined as

$$Q^{\pi}(s_t, u_t) = \mathbb{E}_{\pi}[R_t | s_t, u_t] \quad (15)$$

where the total discounted reward R_t is given by

$$R_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \quad (16)$$

Here, r is the reward obtained at each time step and γ is the discount factor. The discount factor γ determines how much the agent values the reward at the current time step as compared to rewards obtained in the future. The optimal policy π^* is found by using the previous history of the states visited by the agent and the rewards collected by it during its interaction with the environment. By this the agent generates experience which is then used to improve the policy. The action-value function written in a recursive format (also known as the Bellman equation) is given by

$$Q^{\pi}(s_t, u_t) = \mathbb{E}_{r_t, s_{t+1} \sim \mathcal{E}}[r(s_t, u_t) + \gamma \mathbb{E}_{u_{t+1} \sim \pi}[Q^{\pi}(s_{t+1}, u_{t+1})]] \quad (17)$$

where the state s_{t+1} is observed from environment $\mathcal{E}$ due to an action u_t selected from state s_t . It is further assumed that the action u_{t+1} is also selected according to the policy π . Since DDPG uses a deterministic policy and the state transition is deterministic under a selected action in this problem, the above equation then becomes

$$Q^{\mu}(s_t, u_t) = r(s_t, u_t) + \gamma Q^{\mu}(s_{t+1}, \mu(s_{t+1})) \quad (18)$$

where μ represents the deterministic policy function.

DDPG adopts an actor-critic framework where both the policy and action-value functions are learnt using neural networks. The actor network takes in the input the current state of the agent and provides the action to be taken according to the policy as its output. The critic network takes in the action and the state as the inputs and provides the Q value as the output. The goal of the critic network is to minimise the mean square temporal difference (TD) error:

$$L = \frac{1}{N} \sum_{i=1}^N (Q(s_i, u_i | \theta^Q) - y_i)^2 \quad (19)$$

where θ^Q represents the parameters of the critic network, N is the sample batch size, and y_i is the temporal difference (TD) target given by

$$y_i = r(s_i, u_i) + \gamma Q(s_{i+1}, \mu(s_{i+1})) \quad (20)$$

The TD error is defined as difference between the evaluated Q value and the TD target y_i . If y_i is calculated using the same network used by the critic, it may be very hard to converge. So a target critic $Q'(s, u | \theta^{Q'})$ and target actor $\mu'(s | \theta^{\mu'})$ is introduced. The parameters $\theta^{Q'}$ and $\theta^{\mu'}$ is updated using a soft method given by

$$\begin{aligned} \theta^{Q'} &\leftarrow \tau \theta^Q + (1 - \tau) \theta^{Q'} \\ \theta^{\mu'} &\leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'} \end{aligned} \quad (21)$$

where $\tau \ll 1$. So, the TD target y_i can be expressed as

$$y_i = r(s_i, u_i) + \gamma Q'(s_{i+1}, \mu'(s_{i+1} | \theta^{\mu'}) | \theta^{Q'}) \quad (22)$$

The aim of the actor network is to maximise the expected accumulated reward J whose gradient is given by

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_{i=1}^N \nabla_u Q(s, u | \theta^Q) |_{s_i, u=\mu(s_i)} \nabla_{\theta^\mu} \mu(s | \theta^\mu) |_{s_i} \quad (23)$$

where s_i is sampled from the replay buffer $\mathcal{D}$.

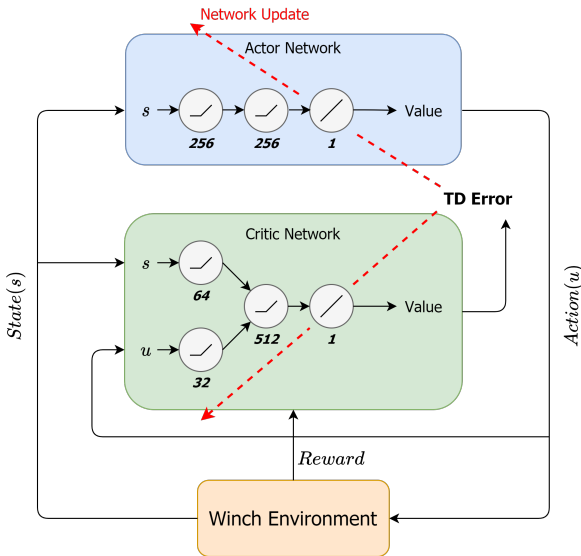


Fig. 4. DDPG Architecture

The RL agent is programmed using Keras with a TensorFlow backend. The training has been performed on Nvidia GeForce GTX 1060 16GB GPU. The state space $s \in \mathbb{S}$ chosen for the application of RL is defined as

$$\mathbb{S} = \{z_w, \dot{z}_w, z_{\text{winch}}, \dot{z}_{\text{winch}}\} \quad (24)$$

Algorithm 1 DDPG algorithm

- 1: Randomly initialise critic network $Q(s, u | \theta^Q)$ and actor $\mu(s | \theta^\mu)$ with weights θ^Q and θ^μ
 - 2: Set target parameters equal to main parameters $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$
 - 3: Empty replay buffer $\mathcal{D}$
 - 4: **for** $episode = 1, M$ **do**
 - 5: Initialise a random noise $\mathcal{N}$ for action exploration
 - 6: Receive initial observation state s_0
 - 7: **for** $t = 1, T$ **do**
 - 8: Select action

$$u_t = \text{clip}(\mu(s_t | \theta^\mu) + \mathcal{N}_t, a_{\text{low}}, a_{\text{high}})$$
 - 9: Execute action u_t in the environment $\mathcal{E}$
 - 10: Observe new state s_{t+1} and reward r_t
 - 11: Store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$
 - 12: Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from $\mathcal{D}$
 - 13: Compute target

$$y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1} | \theta^{\mu'}) | \theta^{Q'})$$
 - 14: Update the critic network by minimising the loss

$$L = \frac{1}{N} \sum_{i=1}^N (Q(s_i, u_i | \theta^Q) - y_i)^2$$
 - 15: Update the actor policy by applying the following gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_{i=1}^N \nabla_u Q(s, u | \theta^Q) |_{s_i, u=\mu(s_i)} \nabla_{\theta^\mu} \mu(s | \theta^\mu) |_{s_i}$$
 - 16: Update the target networks with

$$\begin{aligned} \theta^{Q'} &\leftarrow \tau \theta^Q + (1 - \tau) \theta^{Q'} \\ \theta^{\mu'} &\leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\mu'} \end{aligned}$$
 - 17: **end for**
 - 18: **end for**
-

where z_w is the length of the reeled rope and z_{winch} is the net heave at the winch due to the motion of the vessel in waves. The action space $u \in \mathbb{A}$ is defined as

$$\mathbb{A} = \{u_p\} \quad (25)$$

where u_p is the control input provided to the winch model.

Fig. 4 shows a schematic diagram of the structure of the DDPG architecture used. The pseudo code of the DDPG algorithm is shown above. The actor network is composed of two fully connected layers with 256 neurons each whereas the critic network is composed of a fully connected layers with 64 and 32 neurons for inputs s and u respectively followed by a fully connected layer of 512 neurons as shown in Fig. 4. A rectified linear unit function is used as the activation function for each neuron.

Since the goal in active heave compensation is to minimise the compensated motion without producing any chattering of the payload motion, the reward r is defined as follows

$$r = \begin{cases} 1 - 20e_z - \dot{e}_z & e_z \leq 0.05 \text{ m} \\ -10e_z - 2\dot{e}_z & e_z > 0.05 \text{ m} \end{cases} \quad (26)$$

where $e_z = |z_w + z_{\text{winch}}|$ and $\dot{e}_z = |\dot{z}_w + \dot{z}_{\text{winch}}|$ respectively.

The RL based controller is trained for 150 episodes for 3000 training steps with a sample time of 0.1 sec. The reference used for training is a 300 sec net heave time history at the winch in moderate sea state. During the simulation, the initial conditions of the states are sampled from a uniform distribution. The range of state 1 is $(-1, 1)$, the range of state 2 is $(-10^6, 10^6)$, the range of state 3 is $(-0.1, 0.1)$, and the range of state 4 is $(0, 1)$. The sample batch size is set to 128 with a replay buffer capacity of 50000. For training the network, an Adam optimiser is used for both the actor and the critic network. The learning rate of the actor and the critic network are set to 0.001 and 0.002 respectively. The target network transition gain τ is set as 0.005 and the discount factor γ is set as 0.998. For better exploration, an Ornstein–Uhlenbeck action noise with $\theta = 0.15$, $\mu = 0$, and $\sigma = 0.0005$ is used. Fig 5 shows the learning curve achieved as a result of the training. The average episodic reward in Fig 5 indicates the average of the rewards received in the last 30 episodes.

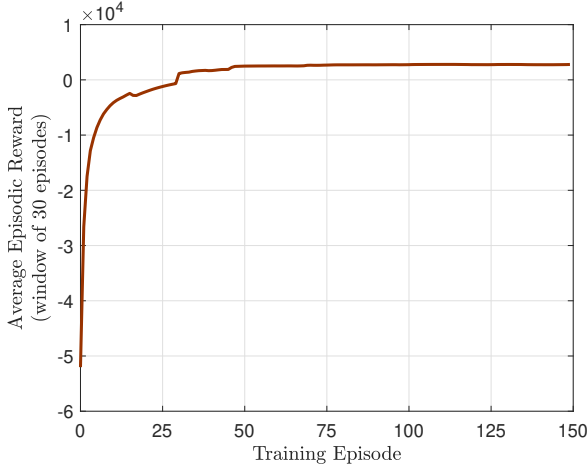


Fig. 5. Learning curve of the RL agent

For comparison, an PD controller is used, with the tuned gains as shown below

$$K_p = 5.86, K_d = 5.46, T_f = 0.03 \quad (27)$$

where K_p , K_d are the controller gains and T_f is the time constant for noise filter. The gains are tuned using loop shaping approach (Zinage and Somayajula (2020)). The control input for this control law in the Laplace domain is given by

$$U(s) = \left(K_p + \frac{K_d s}{1 + T_f s} \right) E(s) \quad (28)$$

where $U(s)$ and $E(s)$ are the Laplace transform of $u_p(t)$ and $e(t) \equiv -(z_w(t) + z_{\text{winch}}(t))$ respectively.

5. SIMULATION RESULTS

The following four cases are analysed to understand the advantages and limitations of using RL based control over classical control.

5.1 Heave compensation with no disturbance and no noise

In this study, heave compensation is defined as the ratio of the difference between the RMS value of uncompensated and compensated net heave at the winch to the RMS value

of uncompensated net heave at the winch. Table 3 shows the comparison of the compensation performance between RL and PD controller for different sea states.

Table 3. Heave compensation in different sea states

Sea State	RL-Control	PD-Control
Slight ($H_s = 1.5$ m, $T_p = 6$ s)	96.9%	86.6%
Moderate ($H_s = 4$ m, $T_p = 9$ s)	99.3%	89.8%
Rough ($H_s = 6$ m, $T_p = 12$ s)	98.1%	92.1%
Very Rough ($H_s = 8.5$ m, $T_p = 14$ s)	95.53%	83.26%

It can be seen that the RL based control is able to demonstrate a better heave compensation performance than the PD control for all the four sea states. As the sea state keeps on increasing the effect of saturation in the swash angle is observed more often thereby leading to a decreased compensation performance in higher sea states. However, as per Table 3 it can be seen that the compensation performance of the RL controller almost remained constant irrespective of sea state. This is indicative that the RL based control is able to handle the saturation in the swash angle better than the PD control. Fig. 6 shows the plot of the uncompensated motion and the compensated motion time histories in rough sea state.

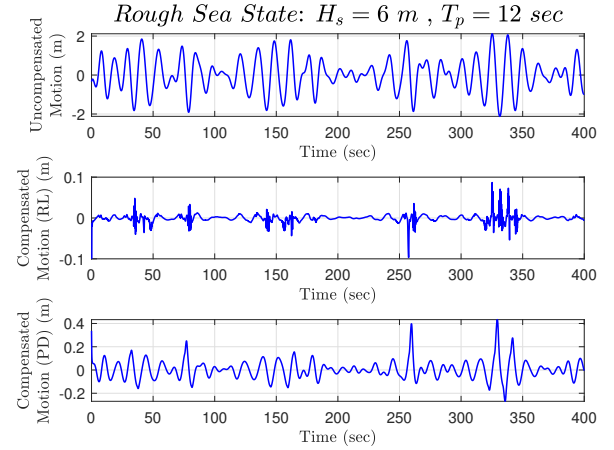


Fig. 6. Uncompensated and compensated motion time histories at the winch for rough sea state.

5.2 Heave compensation with an offset

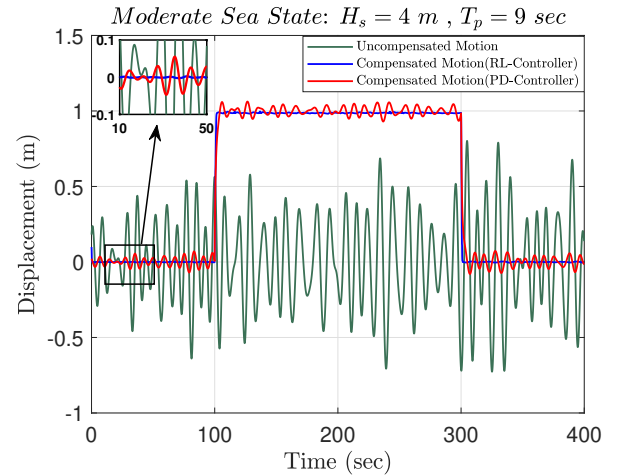


Fig. 7. Ability of the controllers to track the offset

Fig. 7 shows the ability of the both the controllers in tracking an offset for a period of 200 secs for a wave incident angle of 135 degrees. It can be seen that the RL controller performed better than the PD controller while tracking an offset. This type of offset tracking is particularly important in many offshore operations where a load is lowered from an offshore crane onto a floating structure. Topside installation of offshore platforms is an example where offset tracking is an important requirement.

5.3 Heave compensation with disturbance but no noise

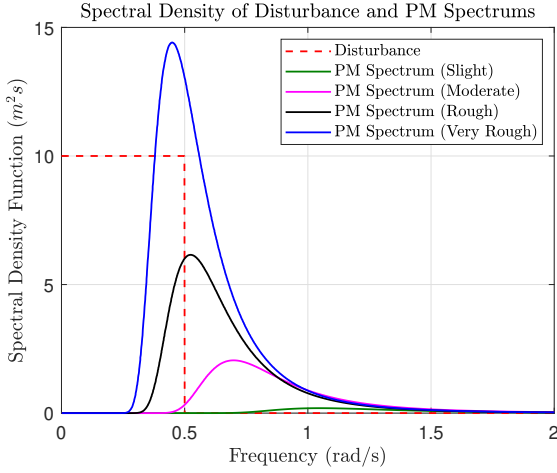


Fig. 8. Spectrum of disturbance and PM spectra's for different sea states

In order to analyze the ability of the controllers to reject disturbance d , a disturbance time history is generated with a constant spectrum of spectral density value 10 with a cut in and cut off frequencies of 0 rad/s and 0.5 rad/s respectively. This disturbance d can be caused due to constant payload torque, unmodelled dynamics, nonlinear friction, vibrations and parametric uncertainty in the real physical system. Fig. 8 shows the plot of the spectrum for the disturbance and the wave elevations in different sea states. In this study, it is assumed that the disturbance entered the system only through the third state (i.e reeled

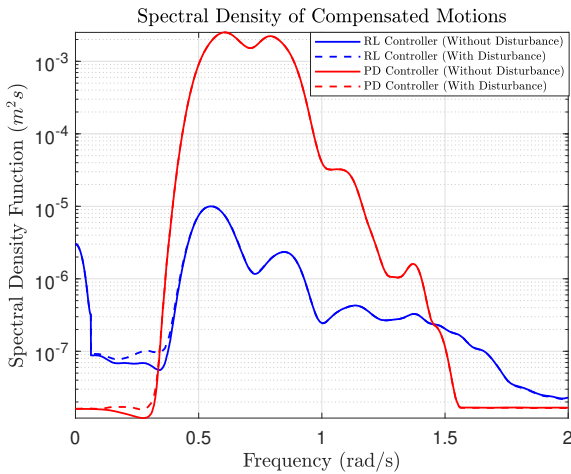


Fig. 9. Spectrum of compensated motions in moderate sea state

rope velocity $\dot{z}_w$). Since the ability of both the control strategies in rejecting disturbances is independent of sea state, the simulation results is analyzed only for the moderate sea state. Fig. 9 shows the power spectral density of the compensated motions for both the controllers in the presence and absence of disturbance. It can be observed that both controllers are good at disturbance rejection.

5.4 Heave compensation with noise but no disturbance

The following two cases are analyzed in this section to test the ability of the controllers in attenuating measurement noise: one with a low noise having spectral density value of 10^{-6} with a cut in and cut off frequency of 3.14 rad/s and 30 rad/s respectively and other with a high noise having spectral density value of 10^{-3} with the same cut in and cut off frequencies. In order to replicate an actual measured output signal, the noise generated is added to z_w and $\dot{z}_w$ and provided as an input to both the controllers.

The signal to noise ratio (SNR) is usually defined in decibels as

$$\text{SNR} = 10 \log_{10} \left(\frac{\sigma_{\text{signal}}^2}{\sigma_{\text{noise}}^2} \right) = 20 \log_{10} \left(\frac{\sigma_{\text{signal}}}{\sigma_{\text{noise}}} \right) \quad (29)$$

In this study, σ_{signal} is taken as the RMS value of the uncompensated net heave motion at the winch and σ_{noise} is taken as the RMS value of the noise. A constant noise spectral density of 10^{-6} between 3.14 rad/s and 30 rad/s corresponds to a SNR value of 34.53 dB whereas a constant noise spectral density of 10^{-3} with the same cut in and

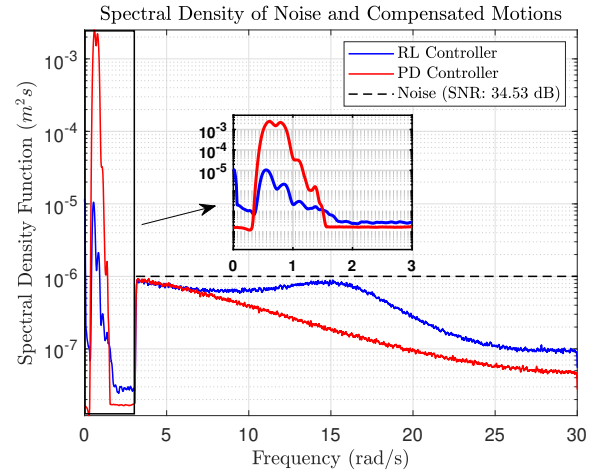


Fig. 10. Spectrum of compensated motions for low noise in moderate sea state

cut off frequencies corresponds to a SNR value of 4.56 dB for moderate sea state.

Fig. 10 shows the power spectral density of the noise and compensated motions for both the strategies when a low noise is included. From Fig. 10, it can be seen that in high SNR case, the PD controller is able to perform better in attenuating the noise at higher frequencies. Also, in this case there is no effect of noise seen on the compensation ability of both the controllers as per Fig. 9 and zoomed plot inside Fig. 10.

Fig. 11 shows the power spectral density of the noise and compensated motions for each of the control strategies

when a high noise is included. As per Fig. 11, it can be seen that RL controller performed better in attenuating higher magnitude noise. Saturation in the normalized swash angle is observed for both the controllers in this case, and due to this the compensation ability is significantly effected. As per Fig. 9 and zoomed plot inside Fig. 11, the PD controller is found to be more affected due to the noise as compared to the RL based controller. When it came to high noise attenuation the RL based controller performed better than the PD controller as shown in Fig. 11.

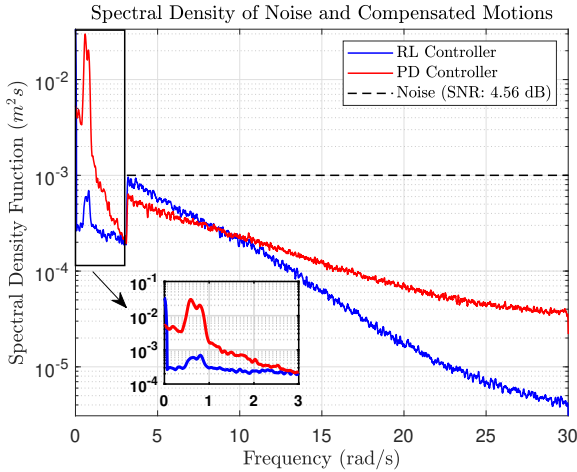


Fig. 11. Spectrum of compensated motions for high noise in moderate sea state

6. CONCLUSION

In this paper, a deep reinforcement learning based controller is proposed for active heave compensation (AHC). A deep deterministic policy gradient (DDPG) algorithm is used to develop a controller for AHC. The control policy is trained by simulating the winch environment for several episodes in a moderate sea state. The extracted control policy is then tested on four different sea states to validate its heave compensation ability. A Proportional-Derivative (PD) controller is tuned and used to compare the results of the RL based controller. It is found that the RL based controller is able to provide a better compensation performance than the PD controller in all four sea states. Both the controllers are reasonably good in disturbance rejection. The PD controller is able to attenuate noise better in a low noise scenario. However, in a high noise scenario, the RL based controller is able to provide better noise attenuation when measurement noise is added to the feedback signal. The RL based controller is also able to handle the saturation dynamics better than the PD controller.

REFERENCES

Do, K.D. and Pan, J. (2008). Nonlinear control of an active heave compensation system. *Ocean engineering*, 35(5-6), 558–571.

Guha, A. (2016). *Development and application of a potential flow computer program: determining first and second order wave forces at zero and forward speed in deep and intermediate water depth*. Ph.d. dissertation, Texas A&M University, College Station, TX.

Hatleskog, J.T. and Dunnigan, M.W. (2007). Passive compensator load variation for deep-water drilling. *IEEE Journal of Oceanic engineering*, 32(3), 593–602.

Konda, V.R. and Tsitsiklis, J.N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, 1008–1014. Citeseer.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1), 1334–1373.

Li, Z., Ma, X., Li, Y., Meng, Q., and Li, J. (2019). Adrc-empc active heave compensation control strategy for offshore cranes. *Ships and Offshore Structures*, 1–9.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Martinsen, A.B. and Lekkas, A.M. (2018). Straight-path following for underactuated marine vessels using deep reinforcement learning. *IFAC-PapersOnLine*, 51(29), 329–334.

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Richter, M., Schaut, S., Walser, D., Schneider, K., and Sawodny, O. (2017). Experimental validation of an active heave compensation system: estimation, prediction and control. *Control Engineering Practice*, 66, 1–12.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, 550(7676), 354–359.

Somayajula, A.S. (2017). *Reliability Assessment of Hull Forms Susceptible to Parametric Roll in Irregular Seas*. Ph.d. dissertation, Texas A&M University, College Station, TX.

Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y., et al. (1999). Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, 1057–1063. Citeseer.

Woo, J., Yu, C., and Kim, N. (2019). Deep reinforcement learning-based controller for path following of an unmanned surface vehicle. *Ocean Engineering*, 183, 155–166.

Woodacre, J., Bauer, R., and Irani, R. (2018). Hydraulic valve-based active-heave compensation using a model-predictive controller with non-linear valve compensations. *Ocean Engineering*, 152, 47–56.

Zhao, L. and Roh, M.I. (2019). Colregs-compliant multiship collision avoidance based on deep reinforcement learning. *Ocean Engineering*, 191, 106436.

Zinage, S. and Somayajula, A. (2020). A comparative study of different active heave compensation approaches. *Ocean Systems Engineering*, 10(4), 373.