Adaptive Inverse Deep Reinforcement Lyapunov learning control for a floating wind turbine

Hadi Mohammadian KhalafAnsar[1,*], Jafar Keighobadi[2],
[1.2] Faculty of Mechanical Engineering, University of Tabriz, East Azerbaijan, Iran.

Abstract Offshore floating wind turbines (FWT) decrease climate change adversial effects without occupying significant land and harvesting fields. Owing to the earth planet unexpected climate, online adaptive feedback control of FWTs will be effective in the sense of optimal and uniform energy capture. In this paper, a deep reinforcement learning (DRL)-based control system is proposed to offset both the disturbance and noise effects. Large variations of wind and water waves generate enormous information give rise to convergent learning of deep neural networks model of the wind turbine. As a result of the disturbance and wind sudden variations, an adaptive inverse control equipped with DRL could easily cope with the inherent drawback of DRL i.e., tracking error. Furthermore, received rewards in the DRL algorithm are passed through the newly designed training algorithm to predict control actions such that the loss function is decreased. The attenuation of disturbance and noise on the tracking performance of closed-loop FWT is clarified through software implementation tests while the weight's convergency and update rules are proved by the direct Lyapunov theorem.
Keywords: Deep reinforcement learning; Artificial intelligence; Adaptive inverse control; Deep deterministic policy gradient; Floating

1. Introduction

Increasing utilization of green and renewable energy guides to online management of floating wind turbine (FWT) for uniform and optimal production. Offshore wind turbines play an important role in green energy production owing to stable and high-speed wind flow in oceans. Besides, conventional control systems are not resistant to the changeable weather conditions in the installed location of FWT. Therefore, design of intelligent control systems is recommended along with efficient confrontation with environmental changes for a stable energy management. In literature of adaptive control, a nonlinear control within self-regulatory procedure has been considered by Widrow and Walach to minimize the disturbance effect [1], [2]. Application of state filters through single-input single-output model series of reference adaptive control and adaptive filtering based on neural networks are respectively proposed in [3], [4]. Regarding feasible results of combined adaptive controllers with artificial neural networks, the stability analysis is carried out. Through training with vast available data, advanced machine learning (ML) and imitation learning get skills from human manifestations [5-7]. One of the significant and mostly used ML approaches is deep reinforcement learning (DRL) which applies two actor-critic networks to adjust Markov decision processes according to the maximization of cumulative rewards. Unlike supervised/unsupervised learning, DRL receives samples without labels and just considers reward functions in determination of selected policy and state-action pairs [8], [9]. The advantage of DRL over the other ML methods is application of minimal prior information to reach optimum control [10-15]. The DRL technique is computationally simple for implementation where it offers a direct method for nonlinear system control by merging capability of optimal and adaptive control algorithms [16], [17]. Consequently, this paper tries to show how adaptive filtering algorithms featuring DRL should be applied to control of time-varying systems subjected to disturbance and exogenous inputs.

Nowadays, as one of the most prominent tools in the energy management field, FWT is developed under the newly proposed feedback controller. In this field, using Colliding Bodies Optimization technique, Kaveh and Sabeti suggested a brand-new approach to implement the FWT system [18]. Mohammadi et.al investigated the new topology for transverse permanent magnet generator for small-scale wind turbine aimed at decrease in issues including unbalanced voltage and high demagnetization [19]. In [20], the authors suggested the enhancement of the surface of flange shrouded wind turbine to optimize the energy harvest. Shamsnia and Parniani compared the performance of the new excitation controlled synchronous generator-based wind turbine with electrically excited wind turbine to show that the proposed new structure is promising in terms of economic, reliability, and efficiency [21]. Abedini et al. investigated the microgrids including power plant and two wind turbines with all possible real-world assumptions as purpose of educational courses [22]. Although the conventional control of FWT has been well studied, the intelligent-based controllers need more research. Some of the ML works in this field are reported as

* Corresponding author. Tel./Fax: 041-33354153; Mobile Number: 09353813756
E-mail addresses: H.MohammadianKhalafAnsar@tabrizu.ac.ir (Hadi Mohammadiyan KhalafAnsar);
keighobadi@tabrizu.ac.ir (Jafar Keighobadi);

follows. Zhang et al. investigated the structural control of FWT employing the active adjustable mass damper and reinforcement learning-based control method [23]. In [24], using an adaptive neuro-fuzzy inference system with a type-2 structure and a passive RL solution by particle swarm optimization policy was suggested to regulate the pitch angle of an actual wind turbine. Bin Tang et al. explained the application of a fuzzy information granulation with an Elman neural network estimation of short-term wind power intervals' fluctuation [25]. Wayne Yao et al. proposed a deep learning method featuring the Long Short-Term Memory (LSTM) model fuzzy-rough set for wind speed prediction [26]. Chengcheng Gu and Hua Li prudently investigated and presented the different approaches and utilizations of deep learning in wind energy [27]. Hui Liu et al. used the Variational Mode Decomposition (VMD) to break the wind speed data into a set, Singular Spectrum Analysis (SSA) to excerpt the learning data of all the set components, LSTM network to fulfill the prediction for the low-frequency components obtained by the VMD-SSA, and Extreme Learning Machine (ELM) to complete the prognostication for the high-frequency components, for producing a unique wind speed multistep forecasting model [28]. Enrique and Santos combined an adaptive neural network, a proportional-integral-derivative (PID), an inverse model of the plant, and two switches to control and track the signals appropriately [29].

A proper control law is required to fulfil numerous standard control protocols especially regarding the structural nonlinear complexity. Up-to-date FWT technology focuses on the structure stabilizing controllers along with being robust in the presence of exogenous inputs/disturbances. In this paper, we develop a 16 degrees of freedom floating turbine model, with three control actions including yaw and pitch angles and generator's torque. Also, we have assumed wave/wind disturbances are formulated as unwanted topographical effects.

Adaptive control methods are designed to deal with uncertainty and noisy signals [30]. In our recent paper [30], we have developed an adaptive dynamic surface control as a class of sliding mode control systems to overcome term explosion in the input action signal. According to the designer experience, suitable filters are common method of stabilizing the control system of adaptive dynamic surface technique. However, in this paper, we propose an adaptive inverse structure, involving both feed-forward and feedback actions, that guarantees noise attenuation together with zero steady-state tracking error. Gathering the deep reinforcement learning inside the self-tuning adaptive inverse control method provides robustness against disturbances and therefore improves the tracking performance upon high noise to signal ratios. Tunning of the DRL weights through direct Lyapunov method enable the lost function extraction while trapping in local minima and alleviates the vanishing gradient problem. Now, new specialized contributions of the current paper are summarized as follows.

1. As a key facility for green energy power generation, the offshore FWT will undergo position and attitude active tracking control by a newly developed adaptive inverse DRL controller.

2. The proposed adaptive inverse technique applies two online DRL networks as feedforward and feedback whichever includes four clarified networks. In the actor-critic of DRLs, two lagged networks are intended to overcome possible divergences of weight learning process.

3. Heuristic development of upper bounded reward functions in the DRL-oriented control of FWT yields in minimal input action leading the turbine's tower operation in the desired position.

4. An improved robust control strategy of DRL against noises is figured out by sensitivity analysis while the environmental condition would change due to modeling uncertainties and possible perturbations.

5. The gradient descent update process of the DRL has been improved through Lyapunov direct method.

This research work aims to study sustainability of the adaptive inverse DRL control system specialized to the FWT system by addressing data efficiency in wide-ranging applications. Consequently, a disturbance model of a single frequency sinusoid profile based on linear wave theory is used in software experiments for performance evaluation purposes. Mostly, the application of exact model to address the disturbance phenomenon comprehensively leads to complex mathematical models which are not straightforward to solve. Moreover, the newly designed controller as a discrete solution of the complex mechanical plants uses black-box simplifying process of dynamical equations.

The rest of this paper is organized as follows. Section 2 discusses the theoretical framework of the DRL algorithm and the proposed experimental procedure. Section 3 explains in detail the nonlinear model of the FWT under consideration and the setup of the simulations. Section 4 reports the numerical results found for the considered environments. Finally, section 5 concludes the whole of paper and obtained results.

2. Research Methodology

Since the combination of DRL and adaptive inverse control is the main contribution of the current research work, the process of both methods is briefly explained. Therefore, the first section is dedicated to explain DRL in deterministic policy gradient.

2.1. Deep deterministic policy gradient (DDPG) training for DRL

DRL as a kind of data-based technique emphasizes on working with Markov decision process. According to Figure 1, the agent receives observation $\boldsymbol{S_t}$ and reward $\boldsymbol{R_t}$ to generate online action $\boldsymbol{A_t}$ for capture of the maximum rewards $\sum R$ during processing time intervals.

The training route of agent uses an off-policy black-box method as the DDPG algorithm [31]. The agent contains two neural networks, i.e., the actor π_φ and critic $\mathbf{Q}_\vartheta$. The network π_φ approximates the action of the observation $\boldsymbol{S_t}$. On the other hand, $\mathbf{Q}_\vartheta$ calculates the Q-value as a criterion to show how much good this action is. Consequently, the main amount of target is computed as follows.

$$y_j = \begin{cases} r_j & \text{terminates at step j+1} \\ r_j + \gamma \max_{a'} Q'_{\vartheta'}\left(s_{j+1}, a; \vartheta'\right) & \text{otherwise} \end{cases} \tag{1}$$

where r_j stands for the reward; γ the discount factor and ϑ' the weights of lagged Q-network. The reward for agent's action is defined as:

$$r = \begin{cases} r_1, & |x| \le x_{lim} \\ r_2, & |x| > x_{lim} \end{cases} \tag{2}$$

where, x_{lim} shows the maximum tolerable fluctuation of the tower during simulation task. r_1 and r_2 are determined as:

$$r_1 = \left(A_r |x|^2 + B_r |\theta|^2 + C_r |u|^2\right) D_r \tag{3}$$

$$r_2 = r_1 + E_r \tag{4}$$

The parameters are given as $A_r = 10^{-1}$, $B_r = 10^{-2}$, $C_r = 50$, $D_r = 10^{-2}$, $E_r = -10^2$. The shape of the reward penalizes the magnitude of $\boldsymbol{x}$, $\boldsymbol{\theta}$, and $\boldsymbol{u}$. Therefore, the optimal action is obtained if the system oscillates with minimum input, and the tower approaches to the desired location at the end of task. By initialization of DRL with random weights, the agent reacts in the environment to gather experiences in a range of dimension D. Each component of the experience range includes four components $(\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{r}, \boldsymbol{S}')$ where $\boldsymbol{S}'$ symbolizes the new state of the system after acting $\mathbf{A_t}$. Accordingly, to obtain the optimum Q-value, the mean square error is minimized:

$$L_{MSE} = \left(y_j - Q\left(s_j, a_j; \vartheta\right)\right)^2 \tag{5}$$

Backpropagation of Eq. (5) with respect to weights ϑ yields updated weights π_φ and the critic network $\mathbf{Q}_\vartheta$. To avoid divergence of backpropagation normally after some epochs, the update algorithm is performed for weights in lagged versions of main networks. In our proposed controller, the update rules are fulfilled by Lyapunov function. This process is carried out in a loop while the ideal policy is obtained.

2.2. Adaptive Inverse Control Implementation

An adaptive control system is designed to cope with the high-dynamic plant and environmental effectcs as Figure 2. Adaptation rules are obtained by direct Lyapunov stability approach. The adaptive inverse control system consists a feedforward section passing through the DRL model. By feedbacking the plant output through the identified DRL and feeding the error signal from the DRL model output, the adaptation rule computes the gains of controller. Furthermore, the adaption rule feeds the DRL based error signal to direct Lyapunov algorithm to find the optimal weights. The adaptation rule is obtained through minimizing the following loss function:

$$\begin{aligned} Loss\ \ Function = & \frac{1}{d}\sum_{i=1}^{d}\left(\mathbf{Q}_{\vartheta}\left(\mathbf{s}_i,\mathbf{a}_i\right)-y_i\left(\mathbf{s}_i,\mathbf{a}_i,r_i,\mathbf{s}_i^{'}\right)\right)^2 \ |_{id} - \\ & \frac{1}{d}\sum_{i=1}^{d}\left(\mathbf{Q}_{\vartheta}\left(\mathbf{s}_i,\mathbf{a}_i\right)-y_i\left(\mathbf{s}_i,\mathbf{a}_i,r_i,\mathbf{s}_i^{'}\right)\right)^2 |_{ff} \\ & +\frac{1}{d}(u-u_e)^2 \end{aligned} \tag{6}$$

with 'id' standing for identification in which the first actor and critic of DRL produce u_e and 'ff' shows the feedforward model of DRL producing u. Minimizing Eq. (6) leads to the updated DRL network and control of the FWT under disturbances.

Theorem 1. In the DRL, the following update of weights from the output layer toward the proior layers results in asymptotically convergence of output tracking error $e(k)$.

$$\begin{aligned} w_{ji}^0(k) &= \frac{1}{nx_i(k)}G_j^1\left(\frac{1}{nw_{ji}^1(k)}G_j^2\left(\frac{\beta^{-k/2}e(k-1)+r/2}{w_{1j}^2(k)}\right)\right) \\ w_{ji}^1(k) &= \frac{1}{nS_j^1(k)}G_j^2\left(\frac{\beta^{-k/2}e(k-1)+r/2}{w_{1j}^2(k)}\right) \\ w_{1j}^2(k) &= \frac{\beta^{-k/2}e(k-1)+r/2}{S_j^2(k)} \\ w_{ji}'^0(k) &= \frac{1}{nx_i(k)}G_j^1\left(\frac{1}{nw_{ji}'^1(k)}G_j^2\left(\frac{1/\gamma\left(\beta^{-k/2}e(k-1)+r/2\right)}{w_{1j}'^2(k)}\right)\right) \\ w_{ji}'^1(k) &= \frac{1}{nS_j'^1(k)}G_j^2\left(\frac{1/\gamma\left(\beta^{-k/2}e(k-1)+r/2\right)}{w_{1j}'^2(k)}\right) \\ w_{1j}'^2(k) &= \frac{1/\gamma\left(\beta^{-k/2}e(k-1)+r/2\right)}{S_j'^2(k)} \end{aligned} \tag{7}$$

***Proof*:** The Bellman function (1), is rewritten as follows:

$$e = Q(k)-r(k)+\gamma Q'(k) \tag{8}$$

Considering a Lyapunov function and its diffrence as:

$$V(k)=\beta^k e^2(k) \tag{9}$$

$$\Delta V(k)=V(k)-V(k-1) \tag{10}$$

$$\Delta V(k)=\beta^k e^2(k)-\beta^{k-1}e^2(k-1) \tag{11}$$

Now, some required definition are released as follows:

$$S_j^1 = F_j^1\left(\sum_{i=1}^{n} w_{ji}^0(k)x_i(k)\right) \tag{12}$$

Where S_j^1 denotes the output of hidden layer 1 and F_j^1 stands for sigmoid activation functions. The corresponding weight coefficients in this layer are denoted by w_{ji}^0.

The same as layer 1, layer 2 is parameterized as:

$$S_j^2 = F_j^2\left(\sum_{i=1}^{n} w_{ji}^1(k) S_i^1\right) \tag{13}$$

Where S_j^2 and F_j^2 represent the output of hidden layer 2 and sigmoid activation function for this layer, respectively. The Q-function as a criterion for how good or bad action was taken by actor is computed as:

$$Q(k) = \sum_{j=1}^{n} w_{1j}^2(k) S_j^2 \tag{14}$$

Where w_{1j}^2 stands for the weight coefficients of output layer. Therefore, replacing the equivalent quantities of the previous layers leads to:

$$Q(k) = \sum_{j=1}^{n} \left\{ w_{1j}^2(k) F_j^2\left[\sum_{i=1}^{n} w_{ji}^1(k) F_j^1\left(\sum_{i=1}^{n} w_{ji}^0(k) x_i(k)\right)\right]\right\} \tag{15}$$

Assuming the lagged form of the network with the primed weights in the same rules of main network yields:

$$Q'(k) = \sum_{j=1}^{n} \left\{ w_{1j}'^2(k) F_j^2\left[\sum_{i=1}^{n} w_{ji}'^1(k) F_j^1\left(\sum_{i=1}^{n} w_{ji}'^0(k) x_i(k)\right)\right]\right\} \tag{16}$$

With considering the update rules of the weights as Eq. (7), where $\mathrm{G}_{\mathrm{j}}^1$ and $\mathrm{G}_{\mathrm{j}}^2$ stand for the reciprocal of sigmoid function, and substituting e of Eq. (8) in Eq. (11) guides to:

$$\Delta V(k) = \beta^k \left(Q(k) - r(k) + \gamma Q'(k)\right)^2 - \beta^{k-1} e^2(k-1) \tag{17}$$

Replacing the Q-function Eq. (15) and lagged version Eq. (16) into Eq. (17) yields:

$$\Delta V(k) = \beta^k \left(\begin{array}{l} \sum_{j=1}^{n} \left\{ w_{1j}^2(k) F_j^2\left[\sum_{i=1}^{n} w_{ji}^1(k) F_j^1\left(\sum_{i=1}^{n} w_{ji}^0(k) x_i(k)\right)\right]\right\} - r(k) \\ +\gamma \sum_{j=1}^{n} \left\{ w_{1j}'^2(k) F_j^2\left[\sum_{i=1}^{n} w_{ji}'^1(k) F_j^1\left(\sum_{i=1}^{n} w_{ji}'^0(k) x_i(k)\right)\right]\right\} \end{array} \right)^2 - \beta^{k-1} e^2(k-1) \tag{18}$$

The update rules introduced in Eq. (7) are imposed on Eq. (18):

$$\Delta V(k) = \beta^k \left(\begin{array}{l} \sum_{j=1}^{n} \left\{ w_{1j}^2(k) F_j^2\left[\sum_{i=1}^{n} w_{ji}^1(k) F_j^1\left(\sum_{i=1}^{n}\left(\frac{1}{n x_i(k)} G_j^1\left(\frac{1}{n w_{ji}^1(k)} G_j^2\left(\frac{\beta^{-k/2} e(k-1) + r/2}{w_{1j}^2(k)}\right)\right)\right) x_i(k)\right)\right]\right\} - r(k) \\ +\gamma \sum_{j=1}^{n} \left\{ w_{1j}'^2(k) F_j^2\left[\sum_{i=1}^{n} w_{ji}'^1(k) F_j^1\left(\sum_{i=1}^{n}\left(\frac{1}{n x_i(k)} G_j^1\left(\frac{1}{n w_{ji}'^1(k)} G_j^2\left(\frac{1/\gamma\left(\beta^{-k/2} e(k-1) + r/2\right)}{w_{1j}'^2(k)}\right)\right)\right) x_i(k)\right)\right]\right\} \end{array} \right)^2 \tag{19}$$

$$-\beta^{k-1} e^2(k-1)$$

Through simplifynig manipulation,

$$\Delta V(k) = \beta^k \left(\begin{array}{l} \sum_{j=1}^{n} \left\{ w_{1j}^2(k) F_j^2\left[\sum_{i=1}^{n} w_{ji}^1(k)\left(\frac{1}{n w_{ji}^1(k)} G_j^2\left(\frac{\beta^{-k/2} e(k-1) + r/2}{w_{1j}^2(k)}\right)\right)\right]\right\} - r(k) \\ +\gamma \sum_{j=1}^{n} \left\{ w_{1j}'^2(k) F_j^2\left[\sum_{i=1}^{n} w_{ji}'^1(k)\left(\frac{1}{n w_{ji}'^1(k)} G_j^2\left(\frac{1/\gamma\left(\beta^{-k/2} e(k-1) + r/2\right)}{w_{1j}'^2(k)}\right)\right)\right]\right\} \end{array} \right) \tag{20}$$

$$-\beta^{k-1} e^2(k-1)$$

Multilpying of terms in Eq. (20) and simplifications leads to:

$$\Delta V(k) = -\left(\beta^{k-1} - 1\right) e^2(k-1) < 0 \tag{21}$$

Therefore, according to Lyapunov stability theory, the tracking error $e(k)$ of updating weights converges asymptotically to zero equilibrium point. Table 1 shows the pseudo code of designed process.

Now, the convergency of DRL in FWT task is investigated. Assuming the state space s_t and the controller's action a_t chosen through selecting different possible actions, the possibility of progress from current state to the next ones is defined as $\mathrm{P}(s_{t+1} \mid s_t, a_t)$ and also according to the cost function, the value of a taken action is measurable. If every policy like the norm of s_t approaching to zero is found, the system will be stabilized and $c(s_t, a_t) = \mathbb{E}_{P(\cdot|s_t,a_t)} s_{t+1}$. The stochastic system stability is guaranteed if $\lim_{t\to\infty} \mathbb{E}_{s_t} c_\pi(s_t) = 0$ for any arbitrarily large initial condition s_0. With $\rho(s_0)$ denoting the distribution of initial states, the transition probability is defined as $P_\pi(s') \triangleq \int_A \pi(a \mid s) P(s' \mid s, a)\, da$. Moreover, the state variation's loop at a specific t as $\mathrm{P}(s \mid \rho, \pi, t)$ is defined iteratively as: $\mathrm{P}(s' \mid \rho, \pi, t+1) = \int_S P_\pi(s' \mid s)\mathrm{P}(s \mid \rho, \pi, t)\, da, \forall t \in Z_+$ and $P(s \mid \rho, \pi, 0) = \rho(s)$. By assuming the Markov chain made by ergodic policy π with a fixed supply $q_\pi(s) = \lim_{t\to\infty} P(s \mid \rho, \pi, t)$, the region of attraction (ROA) is defined as start point for stabilization. The convergency of the trajectory to the equilibrium is satisfied provided that the system starts within the ROA.

Theorem 2: The stochastic system is defined as stable under mean cost definition if a function $L: S \to \mathbb{R}_+$ and positive constants β_1, β_2 and β_3 are available such that,

$$\begin{aligned} \beta_1 c_\pi(s) \le L(s) \le \beta_2 c_\pi(s) \\ \mathbb{E}_{s\sim\mu_\pi}\left(\mathbb{E}_{s'\sim P_\pi} L(s') - L(s)\right) \le -\beta_3 \mathbb{E}_{s\sim\mu_\pi} c_\pi(s) \end{aligned} \tag{22}$$

where,

$$\mu_\pi(s) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N} \mathrm{P}(s_t = s, \rho, \pi, t) \tag{23}$$

is the unlimited distribution.

Proof: If the sample distribution series $\{P(s|\rho, \pi, t), t \in Z_+\}$ converges to $q_\pi(s)$ as t approaches ∞, then based on the Abelian theorem, the set $\left\{\frac{1}{N}\sum_{t=0}^{N} P(\rho, \pi, t), N \in Z_4\right\}$ also converges and $\mu_\pi(a) = q_\pi(s)$. Integrated with the form of μ_π, Eq. (22) concludes that first, on the left-hand-side, $L(s) \le \beta_2 c_\pi(s)$ for all $\beta \in Z$. Since the probability density function $P(s \mid \rho, \pi, t)$ is a limited function over S for all t, thus a coefficient M is available such that

$$P(s \mid \rho, \pi, t) L(s) \le M \beta_2 c_\pi(a), \forall s \in S, \forall t \in Z_+ \tag{24}$$

Second, the series $\left\{\frac{1}{N}\sum_{t=0}^{N} P(s|\rho, \pi, t)\} L(s), N \in Z_+\right\}$ approaches element-wise to the function $q_\pi(s) L(s)$. The Lebesgue's theorem [32] provides the convergency of a set $f_n(s)$ element-wise to a function f and defines with some integrable function g in the sense that,

$$\begin{aligned} &|f_n(s)| \le g(s), \forall s \in \mathcal{S}, \\ &\forall n \lim_{n\to\infty} \int_s f_n(s)\, ds = \int_S \lim_{n\to+\infty} f_n(s)\, \mathrm{d}s \end{aligned} \tag{25}$$

Therefore, the left side of Eq. (22) is written:

$$\begin{aligned}
&\int_{\mathcal{S}} \lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N} P(s|\rho,\pi,t)\left(\int_{\mathcal{S}} P_{\pi}(s'\mid s)\mathrm{L}(s')\mathrm{d}s' - L(s)\right)\mathrm{d}s \\
&= \lim_{N\to\infty} \frac{1}{N}\left(\sum_{t=1}^{N+1} \mathbb{E}_{P(s|\rho,\pi,t)} L(s) - \sum_{t=0}^{N} \mathbb{E}_{P(s|\rho,\pi,t)} L(s)\right) \\
&= \lim_{N\to\infty} \frac{1}{N}\left(\mathbb{E}_{P(s|\rho,\pi,N+1)} L(s) - \mathbb{E}_{\rho(s)} L(s)\right)
\end{aligned} \tag{26}$$

Thus taking the relations above into consideration, Eq. (26) supposes

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N}\left(\mathbb{E}_{P(s|\rho,\pi,N+1)} L(s) - \mathbb{E}_{\rho(s)} L(s)\right) \\
&\le -\beta_3 \lim_{t\to\infty} \mathbb{E}_{P(s|\rho,\pi,t)} c_{\pi}(s)
\end{aligned} \tag{27}$$

Since $\mathbb{E}_{\rho(s)} L(s)$ is a limited quantity and L is positive definite, it yields,

$$\lim_{t\to\infty} \mathbb{E}_{P(s|\rho,\pi,t)} c_{\pi}(s) \le \lim_{N\to\infty} \frac{1}{N}\left(\frac{1}{\beta_3}\mathbb{E}_{\rho(s)} L(s)\right) = 0 \tag{28}$$

Suppose a state $\mathrm{s}_0 \in \left\{\mathrm{s}_0 \mid \mathrm{c}_{\pi}(\mathrm{s}_0) \le \mathrm{b}\right\}$ and a positive d are available such that $\lim_{\mathrm{t}\to\infty} \mathbb{E}_{\mathrm{P}(\mathrm{s}|\mathrm{s}_0,\pi,\mathrm{t})} \mathrm{c}_{\pi}(\mathrm{s}) = \mathrm{d}$ or $\lim_{\mathrm{t}\to\infty} \mathbb{E}_{\mathrm{P}(\mathrm{s}|\mathrm{s}_0,\pi,\mathrm{t})} \mathrm{c}_{\pi}(\mathrm{s}) = \infty$. Since $\rho(\mathrm{s}_0) > 0$ for all initial states in $\left\{\mathrm{s}_0 \mid \mathrm{c}_{\pi}(\mathrm{s}_0) \le \mathrm{b}\right\}$, it follows that $\lim_{\mathrm{t}\to\infty} \mathbb{E}_{\mathrm{s}_{\mathrm{t}}\sim\mathrm{P}(\cdot|\rho,\pi,\mathrm{t})} \mathrm{c}_{\pi}(\mathrm{s}_{\mathrm{t}}) > 0,$ which is inconsistent with Eq. (28). Therefore, being in $\forall \mathrm{s}_0 \in \left\{\mathrm{s}_0 \mid \mathrm{c}_{\pi}(\mathrm{s}_0) \le \mathrm{b}\right\}$ leads to $\lim_{\mathrm{t}\to\infty} \mathbb{E}_{\mathrm{P}(\mathrm{s}|\mathrm{s}_0,\pi,\mathrm{t})} \mathrm{c}_{\pi}(\mathrm{s}) = 0.$ Thus the system is stable in mean cost.

2.3. Dynamic Model Description

Among applications of DRL in literature, leakage in the field of the FWT is observed. Hence, an exact model of the FWT is developed through DRL for adaptive control purposes in the presence of disturbances.

Figure 3 represents the wind turbine model, consisting of the aerodynamic force $\vec{F}_A$. buoyancy force $\vec{F}_B$. catenary line forces $\vec{F}_C$ and hydrodynamic drag/inertial force $\vec{F}_D$. For each of these forces, an associated torque is considered as $\vec{T}_A, \vec{T}_B, \vec{T}_C$ and $\vec{T}_D$.

The simulation block-diagram of wind turbine and the corresponding contol system based on DRL are depicted in Figure 4. The learning dynamical model under persistant forces is appropiately applied in parametric identification and sensitivity analysis of designed DRL control method. Considering the vectors of states x, control inputs u including yaw and pitch angles and generator's torque, and showing disturbances with v and w. the equations of motion are considered as described in [30]:

$$\dot{x} = f(x,u,v,w) = \begin{bmatrix}
\dot{\vec{x}}_g \\
\dot{\vec{\theta}}_g \\
\left(\omega_r - \frac{1}{N_{GR}}\omega_g\right) \\
\left(m_g I_{3\times 3} + diag\left[\vec{m}_a\right]\right)^{-1} \sum\left(\vec{F}_A + \vec{F}_B + \vec{F}_C + \vec{F}_D\right) \\
\left(\mathrm{R}\mathrm{I}_g^{-1}\mathrm{R}^T\right)\sum\left(\vec{T}_A + \vec{T}_B + \vec{T}_C + \vec{T}_D\right) \\
\sum_{k_r} \frac{1}{J_r} Q_{k_r}(x,u,v) \\
\sum_{k_g} \frac{1}{J_g} Q_{k_g}(x,u,v)
\end{bmatrix} \tag{29}$$

with

$$
\begin{aligned}
\vec{F}_A &= \frac{1}{2}\rho A_r C_t(\lambda,\beta)\|\vec{v}_n\|\vec{v}_n,\\
\vec{F}_B &= \rho_\omega g A_1 l_i\ \hat{e}_3,\\
\vec{F}_C &= \begin{pmatrix} F_x(\vec{x}_t)\,\mathrm{proj}_{\hat{x}}(\vec{x}_t)\\ F_x(\vec{x}_t)\,\mathrm{proj}_{\hat{y}}(\vec{x}_t)\\ F_y(\vec{x}_t)\end{pmatrix},\\
\vec{F}_D &= K_d\|\vec{v}_t\| + K_a\vec{a}_t,\\
P &= \frac{1}{2}\rho A_r C_p(\lambda,\beta)\|\vec{v}_n\|^3,\\
\vec{\omega}_r &= \frac{1}{J_r}\left(\frac{P}{\omega_r} - k_r\left(\theta_r - \frac{1}{N_{gr}}\theta_g\right) - b_r\left(\omega_r - \frac{1}{N_{gr}}\omega_g\right)\right)\\
\vec{\omega}_g &= \frac{1}{J_g}\left(-T_g + \frac{k_r}{N_{gr}} + \frac{b_r}{N_{gr}}\left(\omega_r - \frac{1}{N_{gr}}\omega_g\right)\right)
\end{aligned}
\tag{30}
$$

In Table 2, the attributes of Eqs. (29) and (30) are fully provided.

To simulate the system with purpose of getting maximum amount of energy, putting the nacelle of turbine in the direction of blowing wind is required. However, the restrictions in selection of the actuator cause to apply of the mean of the trajectory to a smooth and easy-to-implement input action. Second, since the extremum of power is considered, it is essential the derivative of power with respect to the effective parameters to be zero. From mathematical point of view, setting the instant alteration in power to zero yields:

$$\delta P(t) = \frac{\partial P(t)}{\partial x}\delta x + \frac{\partial P(t)}{\partial u}\delta u + \frac{\partial P(t)}{\partial v}\delta v = 0 \tag{31}$$

where the power $P(t)$ is obtained from Eq. (30). If we assume the variation of power with repect to change in states is negligible, Eq. (31) by inserting the relevant inputs leads to:

$$\frac{\partial P(t)}{\partial \beta}\delta\beta + \frac{\partial P(t)}{\partial \gamma}\delta\gamma + \frac{\partial P(t)}{\partial v}\delta v = 0 \tag{32}$$

Therefore, the variational quantity of β angle namely pitch as the control goal is:

$$\delta\beta = -\left(\frac{\partial P(t)}{\partial \gamma}\delta\gamma + \frac{\partial P(t)}{\partial v}\delta v\right)\Big/\left(\frac{\partial P(t)}{\partial \beta}\right) \tag{33}$$

Third, the generator torque's adjustment, is carried out regarding variations in generator speed. In practice, the principle for regulating the generator torque is to keep the rotor speed constant:

$$\dot{\omega}_r = \frac{1}{J_r}\left(\frac{P}{\omega_r} - (N_{GR})T_g\right) = 0 \tag{34}$$

which follows:

$$\delta T_g = \frac{1}{N_{GR}\omega_r}\left(\frac{\partial P(t)}{\partial \beta}\delta\beta + \frac{\partial P(t)}{\partial \gamma}\delta\gamma + \frac{\partial P(t)}{\partial v}\delta v\right) \tag{35}$$

Solving Eq. (35) produces the generator torque required to exploit the uniform energy.

The simulation of the FWT in MATLAB is illustrated in Figure 5, and the oscillation of the outputs through the application of the above-mentioned controller is apparent.

Moreover, the produced power resuting from this system is reported as shown in Figure 6. As it is shown, the average amount of generated power during this operation is $1.6\times10^{7}\,W$ and it allows the consumer ease of access without large fluctuations in energy harvest.

3. Numerical Results

The proposed neural network, therein weights and input-output variables are revealed as figure 7. For actor and critic networks in DDPG, the same inputs are considered. However, the mixing addition layer correspondes to the second layer of critic and the last layer of actor stands for the difference between these two networks.

The training process of the agent is done by employing the DDPG algorithm under the following restrictions:

$$\mathrm{u} \le u_{max}, \qquad |x| \le x_{lim} \qquad t \le t_f \tag{36}$$

the maximum amount of control action u_{max}, is set to $[18rad, 4\times10^{4}\,N.m, 20rad]$ for roll angle, generator torque and pitch angle, respectively; x_{lim} is the maximum lateral output of the plant, and the maximum duration t_f of the task is set to 1000 epochs or more. Then, the sensitivity of the control system comprises of its estimated act under altered environments like modifications with respect to huge exagenous input data and noise is assessed.

According to the impact of modifying environmental conditions on the efficiency of the FWTs designed in this work, the sustainability of the adaptive inverse DRL control system has been concluded. Therefore, the simulation results illustrate the efficacy of the transfer learning methods, especially in the experiment fields such as dynamical systems. Owing to the complexity of the system, nonlinear models do not take into account disturbances [33–37]. In contrast, a back-box method, especially the proposed approach in this paper allows us to consider the impact of this harmful phenomenon. In addition, owing to the adaptation capacity of neural networks, the robustification properties of the system to the uncertainty is high, as shown in Figures 8 through 11 considering adaptive inverse deep reinforcement learning as AIDRL. Moreover, the reliability of the method due to the application of the Lyapunov stability for updating weights is high.

For comparison purposes, pseudorandom binary sequence was added to the system output measurements. The results were shown in Figures 8 and 11 for the newly designed controller through DDPG training and tracking of reference trajectory is vivid. The reason for the superiority of DDPG is the randomness of chosen actions among various ones and it takes a short time for the controller to reach the rest point. Moreover, the required time in the initialization of the controller is due to the identification and adaptation process. On the other hand, this short time to adaptation in comparison with conventional artificial intelligent networks controllers is insignificant and illustrates the superiority of the proposed controller in the field of green and renewable energy.

A sensitivity analysis is performed to evaluate the impact of important system parameters on the state and output variables of the considered application, as well as to design a novel control procedure. Consider a model of finite-dimensional dynamical system described by ordinary differential equations,

$$\dot{x}(t) = f\left(x(t), p, u(t)\right) \tag{37}$$

where $x \in R^n$ is the state vector, $p \in R^m$ stands for the parameter vector and u control inputs are constant for each time interval $t \in [tk;\ tk+1), 0 \le tk < tk+1.$

By solving the differential equations, the sensitivity of the solution $x(t)$ of (37) with regard to a time-invariant parameter vector p is calculated.

$$\dot{s}_i(t) = \frac{\partial f\left(x(t),\ p, u(t)\right)}{\partial x} \cdot s_i(t) + \frac{\partial f\left(x(t), p, u(t)\right)}{\partial p_i} \tag{38}$$

with

$$s_i(t) = \frac{\partial x(t)}{\partial p_i} \in R^n$$

for all $i = 1, \ldots, m$ with the corresponding initial values at $t = t_k$

$$s_i(t_k) = \frac{\partial x(t_k, p)}{\partial p_i} \tag{39}$$

The sensitivity Eqs. (38) need not have to be developed in symbolic form, as discussed in [38], [39]. Instead, it is sufficient to define the dynamic model Eq. (37) and obtain Eqs. (38) using an algorithmic differentiation toolbox. All partial derivatives required in Eq. (38) are calculated using such a toolbox via operator overloading, for example,

with Python software. The system states $x(t)$ must then be assessed along their trajectories for $t \geq t_k$. Through wide range of variation owing to wind, sensitivity analysis in Figures 12 to 14 show the minority changes in states after usage of adaptive inverse DRL controller.

4. Conclusion

In this paper, we carried out modeling and regulation of a FWT system exposed to varying environment and actuating force-torques. Hence, the stabilization of the turbine structure and the exploitation of uniform energy fall within the scope of the authors' research domain. The mentioned scope needs more study on the controllers' structure where the newly designed controller leads to satisfying results in the green energy field.

Besides, the sensitivity analysis of the DRL controller applied to the FWT system was performed. Via great simulated experiments, the reliability of the designed controller was examined. The highest uniform limit of estimation error for state vector and weighted parameter models is guaranteed by Lyapunov's direct method. To further prove the robustness of the designed controller, the future works would take into account the studying of randomly changed environment conditions. Therefore, the present study focused on the computational elements of the FWT system stabilization situation. Through the designed procedures, a nonlinear controller using DRL was proposed and modeled in a virtual environment. The numerical results of the controller are obtained in Matlab, showing that the desired characteristics of the FWT are captured. Further extension of this work may investigate the establishment of a deeper RL-based sensitivity and the tracking accuracy of the control system in a real time experimental test.

Acknowledgments

This work has no funding resources.

Credit author statement

Hadi Mohammadiyan KhalafAnsar: Conceptualization, Methodology, Data testing, Writing – original draft. Jafar Keighobadi: Visualization, Writing – reviewing and editing.

Technical biography

Hadi Mohammadiyan KhalafAnsar received his M.Sc. degree in Mechanical Engineering from University of Tabriz, Iran. Since 2020, he has been working toward the Ph.D. degree with the Department of Mechanical Engineering, University of Tabriz. His research interests include Integrated deep learning, Deep reinforcement learning, Adaptive control, Neuro-fuzzy controller, and Floating wind turbine control.

Dr. Jafar Keighobadi received the Ph.D. degree in Mechanical Engineering and Control Systems from Amirkabir University of Technology, Iran, in 2008. He is currently the Professor of Mechanical Engineering Department at University of Tabriz. His research interests include Artificial intelligence, Estimation and identification, Nonlinear robust control, and GNC.

Reference

1. Widrow, B., Duvall, K.M., Gooch, R.P., and Newman, W.C., "Signal cancellation phenomena in adaptive antennas: Cases and cures", *IEEE Trans, Antennas Propag*, 30, pp.469-478 (1982).

2. Jim, C.W., "A comparison of two LMS constrained optimal array structures", *Proc IEEE*, 65, pp. 1730-1731 (1977).

3. Griffiths, L.J.; Jim, C.W., "An alternative approach to linearly constrained adaptive beamforming", *IEEE Trans, Antenna Propag*, 30, pp. 27-34 (1982).

4. Gooch, R.P., "Adaptive pole-zero array processing", *in Proc, 16th Asilomar Conf, Circuits Syst, Comput, Santa Clara, CA* (1982).

5. Hesse, M., Timmermann, J., Hullermeier, E., and Trachtler, A., "A Reinforcement Learning Strategy for the Swing-Up of the Double Pendulum on a Cart", *Procedia Manuf* (2018).

6. Manrique, C., Pappalardo, C. M., and Guida, D., "A model validating technique for the kinematic study of two-wheeled vehicles", *Springer International Publishing, Odessa, Ukraine,* pp. 549-558 (2020).

7. Pappalardo, C.M., De Simone, M.C., and Guida, D., "Multibody modeling and nonlinear control of the pantograph/catenary system", *Arch. Appl, Mech*, 89, pp. 1589-1626 (2019).

8. Pappalardo, C., and Guida, D., "Forward and Inverse Dynamics of a Unicycle-Like Mobile Robot", *Machines*, 7(1), (2019).

9. Villecco, F., and Pellegrino, A., "Evaluation of Uncertainties in the Design Process of Complex Mechanical Systems", *Entropy*, 19(9), pp. 475 (2017).

10. Villecco, F., and Pellegrino, A., "Entropic Measure of Epistemic Uncertainties in Multibody System Models by Axiomatic Design", *Entropy*, 19, pp. 291 (2017).

11. Hu, D., Pei, Z., and Tang, Z., "Single-Parameter-Tuned Attitude Control for Quadrotor with Unknown Disturbance", *Appl, Sci.,* 10, pp. 5564 (2020).

12. Talamini, J., Bartoli, A., De Lorenzo, A.D., and Medvet, E., "On the Impact of the Rules on Autonomous Drive Learning", *Appl. Sci*, 10, pp. 2394 (2020).

13. Sharifzadeh, S., Chiotellis, I., Triebel, R., and Cremers, D. "Learning to drive using inverse reinforcement learning and deep q-networks", *arXiv:1612.03653* (2016).

14. Cho, N.J., Lee, S.H., Kim, J.B., and Suh, I.H., "Learning, Improving, and Generalizing Motor Skills for the Peg-in-Hole Tasks Based on Imitation Learning and Self-Learning", *Appl. Sci*, 10, pp. 2719 (2020).

15. Zhang, H., Qu, C., Zhang, J., and Li, J., "Self-Adaptive Priority Correction for Prioritized Experience Replay", *Appl. Sci*, 10, pp. 6925 (2020).

16. Hong, D., Kim, M., and Park, S., "Study on Reinforcement Learning-Based Missile Guidance Law", *Appl. Sci*, 10, pp. 6567 (2020).

17. Rivera, Z.B., De Simone, M.C., and Guida, D., "Unmanned Ground Vehicle Modelling in Gazebo/ROS-Based Environments", *Machines*, 7, pp. 42 (2019).

18. Kaveh, A., and S. Sabeti., "Optimal design of monopile offshore wind turbine structures using CBO, ECBO, and VPS algorithms", *Scientia Iranica*, 26(3), pp. 1232-1248 (2019).

19. Mohammadi Ajamloo, A., K. Abbaszadeh, and R. Nasiri-Zarandi., "A novel transverse flux permanent magnet generator for small-scale direct drive wind turbine application: Design and analysis", *Scientia Iranica* 28(6), pp. 3363-3378 (2021).

20. Maftouni, N., and M. Taghaddosi, "A CFD study of a flanged shrouded wind turbine: Effects of different flange surface types on output power", *Scientia Iranica*, 29(1), pp. 101-108 (2022).

21. Shamsnia, Ali, and Mostafa Parniani, "A comparative analysis of the new excitation controlled synchronous generator-based wind turbine", *Scientia Iranica*, 29(1), pp. 151-167 (2022).

22. Abedini, Moein, et al. "Smart microgrid educational laboratory: An integrated electric and communications infrastructure platform", Scientia Iranica 29.5: 2552-2565 (2022).

23. J. Zhang, X. Zhao, and X. Wei, "Reinforcement Learning-Based Structural Control of Floating Wind Turbines", *in IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(3) , pp. 1603-1613 (2022).

24. Tang, B., Chen, Y., Chen, Q., and Su, M. "Research on short-term wind power forecasting by data mining on historical wind resource", *Applied Sciences*, 10(4) (2020).

25. Yao, W., Huang, P., and Jia, Z., "Multidimensional LSTM networks to predict wind speed", *In 2018 37th Chinese Control Conference (CCC)*, pp. 7493-7497 (2018).

26. Gu, C., and Li, H., "Review on Deep Learning Research and Applications in Wind and Wave Energy", *Energies*, 15(4) (2022).

27. Liu, H., Mi, X., and Li, Y., "Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM", *Energy Conversion and Management*, 159, pp. 54-64 (2018).

28. Sierra-García, J. E., and Santos, M. "Switched learning adaptive neuro-control strategy", *Neurocomputing*, 452, pp. 450-464 (2021).

29. Fausto Pedro García Márquez, and Ana María Peco Chacón, "A review of non-destructive testing on wind turbines blades", *Renewable Energy*, 161, pp. 998-1010 (2020).

30. Keighobadi, J.. KhalafAnsar, H. M., and Naseradinmousavi, P., "Adaptive neural dynamic surface control for uniform energy exploitation of floating wind turbine", *Applied Energy*, 316, pp. 119132 (2022).

31. Manrique Escobar, Camilo A., Carmine M, Pappalardo, and Domenico Guida. "A Parametric Study of a Deep Reinforcement Learning Control System Applied to the Swing-Up Problem of the Cart-Pole" *Applied Sciences*, 10(24), pp. 9013 (2020).

32. Van der Vaart, H. R.. and Yen, E. H. "Weak Sufficient Conditions for Fatou's Lemma and Lebesgue's Dominated Convergence Theorem", *Mathematics Magazine*, 41(3), pp. 109-117, (1968).

33. Frarn, ois-Lavet, V., Henderson, P., Islam, R., Bellemare, M.G., and Pineau, J. "An Introduction to Deep Reinforcement Learning", *Found. Trends Mach, Learn*, 11, pp. 219-354 (2018).

34. Yang, Y., Li, X., and Zhang, L., "Task-specific pre-learning to improve the convergence of reinforcement learning based on a deep neural network", *In Proceedings of the 2016 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China*, 2016, pp. 2209-2214 (2016).

35. Zagal, J.C., Ruiz-del-Solar, J., and Vallejos, P., "Back to reality: Crossing the reality gap in evolutionary robotics", *IFAC Proc*, 2004(37), pp. 834-839 (2004).

36. Bekar, C., Yuksek, B., and Inalhan, G., "High Fidelity Progressive Reinforcement Learning for Agile Maneuvering UAVs", *In Proceedings of the AIAA Scitech 2020 Forum; American Institute of Aeronautics and Astronautics, Orlando, FL. USA*, pp. 1-12 (2020).

37. Al-Araji, A.S., "An adaptive swing-up sliding mode controller design for a real inverted pendulum system based on Culture-Bees algorithm", *Eur. J, Control* 2019(45), pp. 45-56 (2019).

38. A. Rauh, V. Grigoryev, H. Aschemann, and M. Paschen, "Incremental Gain Scheduling and Sensitivity-Based Control for Underactuated Ships," in *Proc. of IFAC Conference on Control Applications in Marine Systems, CAMS 2010, Rostock-Warnemunde, Germany*, (2010).

39. A. Rauh and H. Aschemann, "Sensitivity-Based Feedforward and Feedback Control Using Algorithmic Differentiation", in *Proc. of IEEE Intl. Conference on Methods and Models in Automation and Robotics MMAR 2010, Miedzyzdroje, Poland*, (2010).

List of figures

List of tables

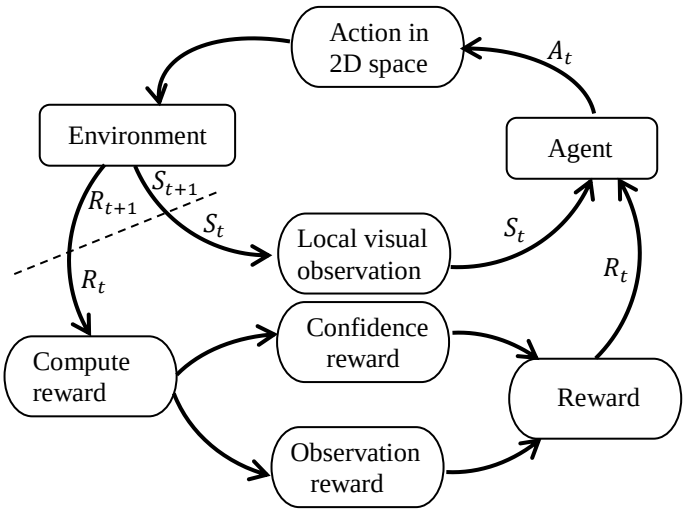

Figure 1

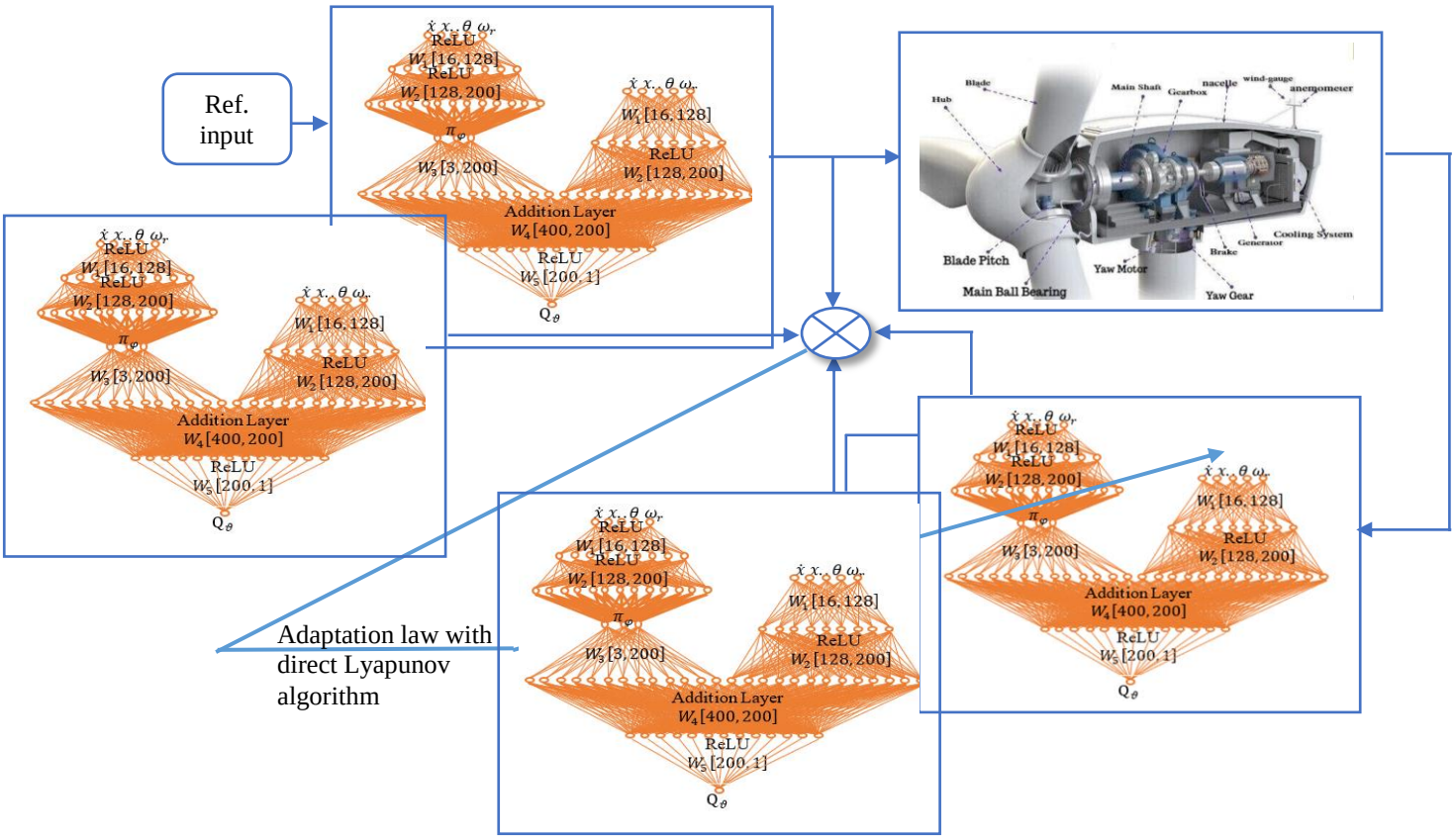
Ref. input
Adaptation law with direct Lyapunov algorithm
Addition Layer
Blade
Hub
Main Shaft
Gearbox
nacelle
wind gauge
anemometer
Blade Pitch
Yaw Motor
Main Ball Bearing
Yaw Gear
Brake
Generator
Cooling System

Figure 2

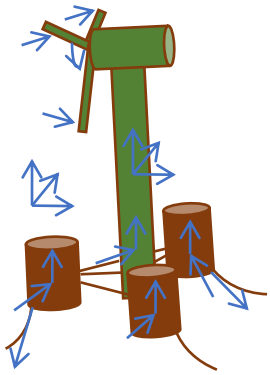

Figure 3

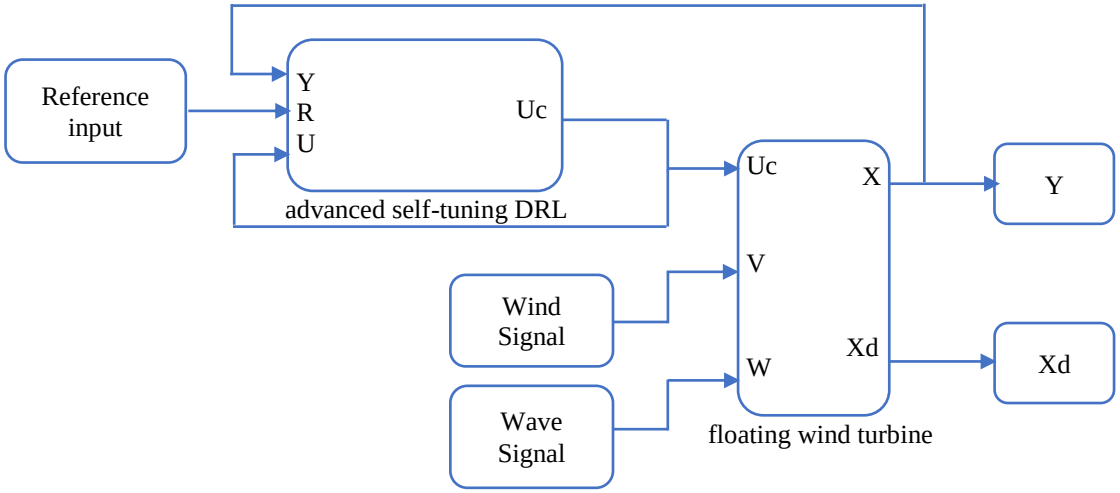

Figure 4

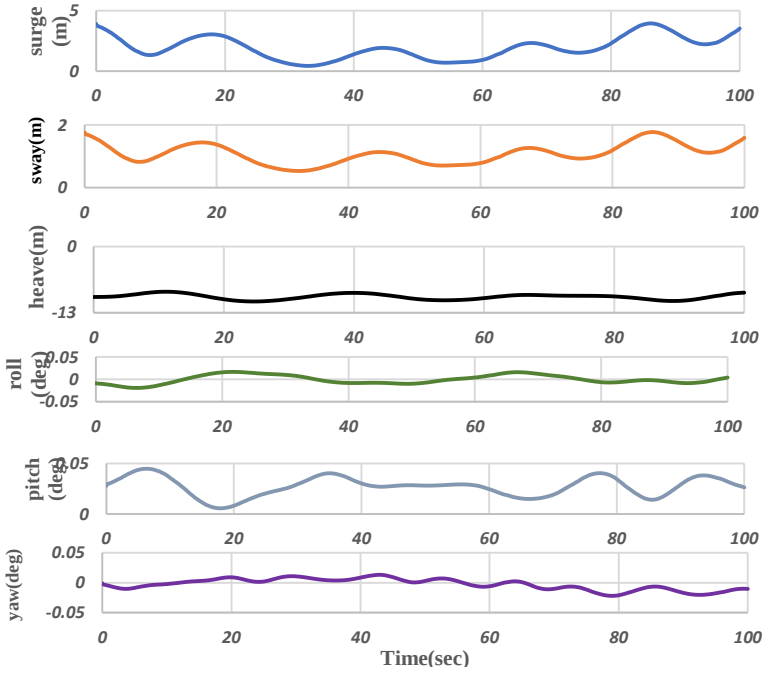

Figure 5

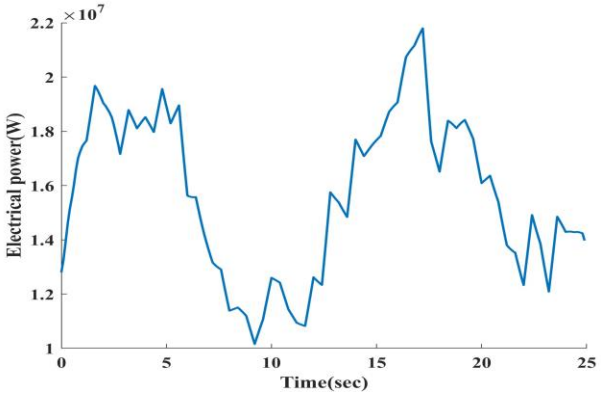

Figure 6

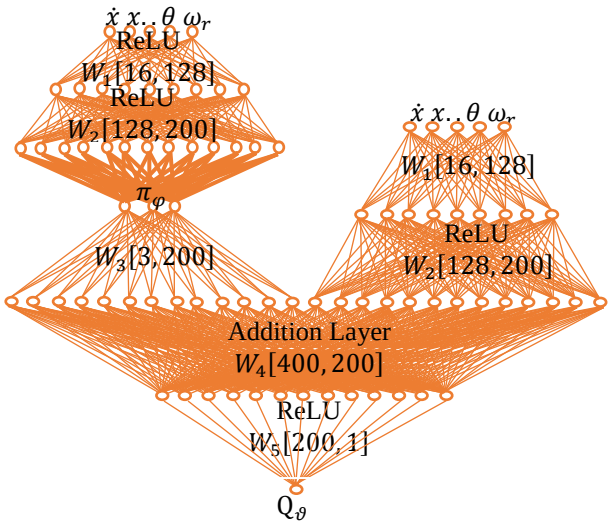

Figure 7

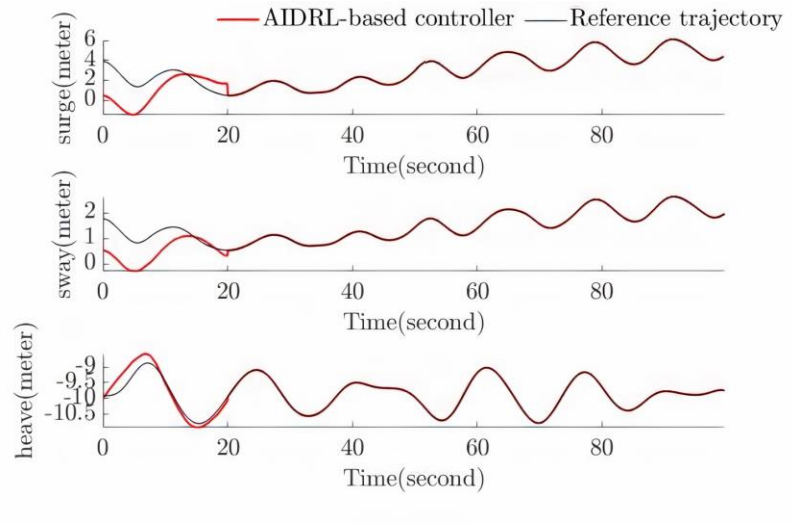

Figure 8

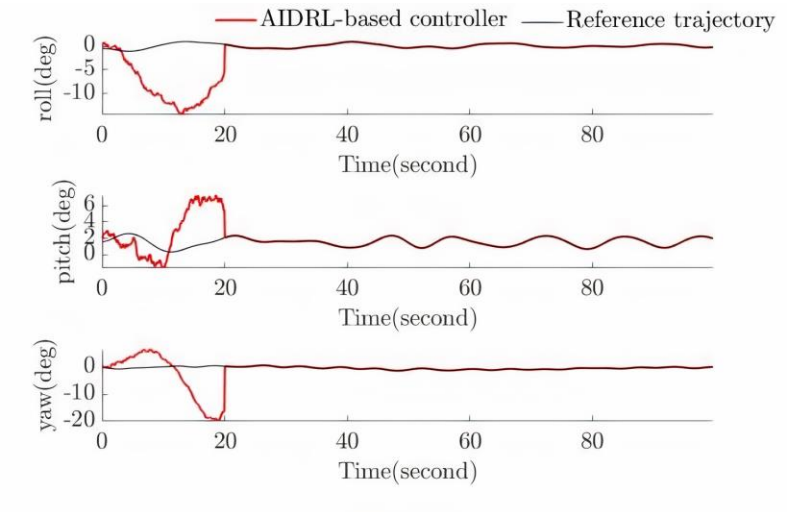

Figure 9

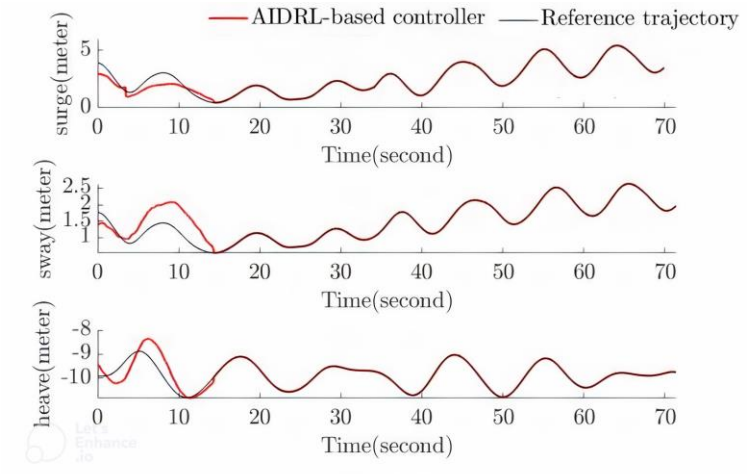

Figure 10

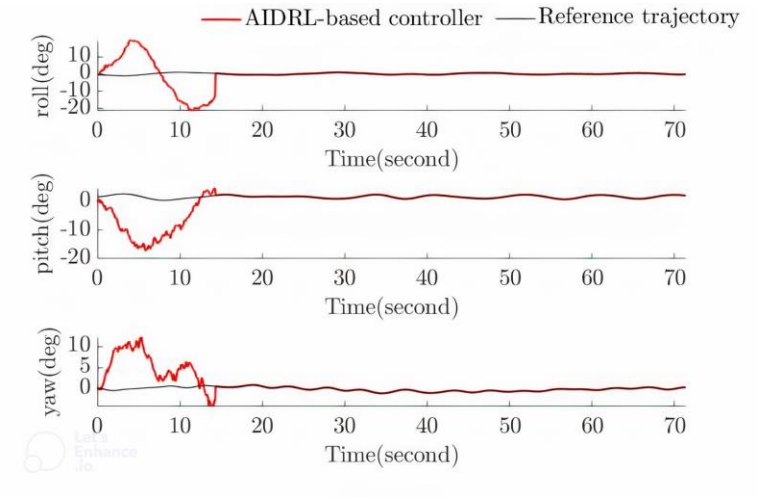

Figure 11

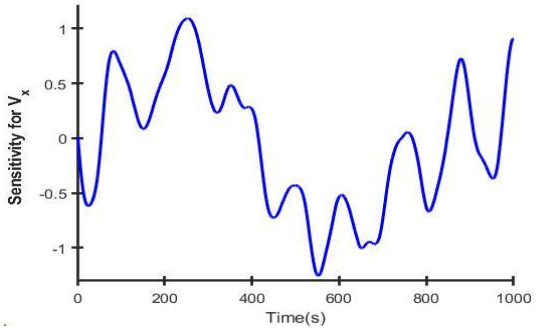

Figure 12

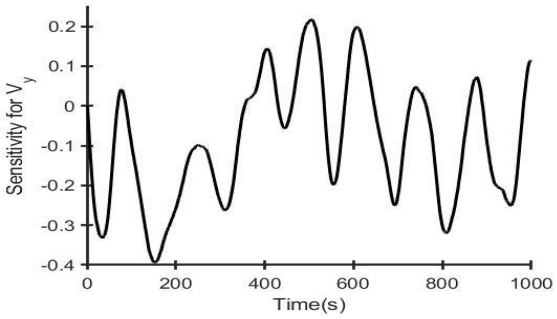

Figure 13

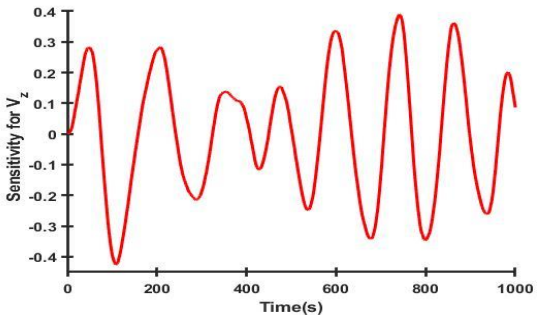

Figure 14

Table 1

Algorithm 1: DDPG algorithm

Randomly initialization of weights ϑ and φ in critic network $\mathbf{Q}_{\vartheta}\left(s,a\,|\,\vartheta\right)$ and actor $\pi_{\varphi}\left(s\,|\,\varphi\right)$.

Initialization of target lagged network $\mathbf{Q}_{\vartheta'}^{'}$ and $\pi_{\phi'}^{'}$ with weights $\vartheta' \leftarrow \vartheta, \phi' \leftarrow \varphi$

Replay buffer *R* Initialization
for episode $=1:M$ do
  Receive initial observation state s_1
  for $\mathrm{t}=1:\mathrm{T}$ do
    Select action $a_t=\mu\left(s_t\,|\,\theta^{\mu}\right)$ according to the current policy
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition $\left(s_t,a_t,r_t,s_{t+1}\right)$ in R
    Sample a random minibatch of *N* transitions $\left(s_i,a_i,r_i,s_{i+1}\right)$ from R
    Set $y_i=r_i+\gamma\mathbf{Q}_{\vartheta'}^{'}\left(\mathbf{s}_i^{'},\pi_{\phi'}^{'}\left(\mathbf{s}_i^{'}\right)\right)$
    Update critic by differentiation of the loss: $L=\frac{1}{N}\sum\left(\mathbf{Q}_{\vartheta}\left(\mathbf{s}_i,\mathbf{a}_i\right)-y_i\left(\boldsymbol{s}_i,\,\boldsymbol{a}_i,\,\boldsymbol{r}_i,\,\mathbf{s}_i^{'}\right)\right)^2$ with respect to weights
    Update the actor policy using the sampled policy gradient:

$$\begin{aligned} Loss\ \ Function = \ & \frac{1}{d}\sum_{i=1}^{d}\left(\boldsymbol{Q}_{\vartheta}\left(\boldsymbol{s}_i,\boldsymbol{a}_i\right)-y_i\left(\boldsymbol{s}_i,\boldsymbol{a}_i,\boldsymbol{r}_i,\boldsymbol{s}_i^{'}\right)\right)^2\ |_{id} - \\ & \frac{1}{d}\sum_{i=1}^{d}\left(\boldsymbol{Q}_{\vartheta}\left(\boldsymbol{s}_i,\boldsymbol{a}_i\right)-y_i\left(\boldsymbol{s}_i,\boldsymbol{a}_i,\boldsymbol{r}_i,\boldsymbol{s}_i^{'}\right)\right)^2|_{ff} \\ & +\frac{1}{d}(u-u_e)^2 \end{aligned}$$

    Update the target networks using Eq. (8)

  end for
end for

Table 2

Property	Variable	Value	Unit
Total inertia about x-axis of body frame	I_{xx}	1.695e10	$kg\cdot m^2$
Total inertia about y-axis of body frame	I_{yy}	1.695e10	$kg\cdot m^2$

Total inertia about z-axis of body frame	I_{zz}	1.845e10	$kg \cdot m^2$
Physical mass	m_g	14.072.718	kg
Air density	ρ_a	1.225	kg / m^3
Effective rotor radius	R_r	62.94	m
drift coefficient	C_t	-	-
tip speed ratio	λ	-	-
blade pitch angle	β	-	deg
normal velocity onto the surface of the rotor blades	$\vec{v}_n$	-	m/sec
Water density	ρ_ω	1025	kg / m^3
gravity constant	g	9.806	m / sec^2
Diameter of columns	D_i	24	m
length of cylinder	l_i	6	m
total 3-D distance vector between the anchor point of the mooring line and the attachment point on the turbine	$\vec{x}_t$	-	m
drag constant of the Morrison equation	K_d	-	-
inertia constant of the Morrison equation	K_a	-	-
transverse velocities	$\vec{v}_t$	-	m / sec
transverse accelerations	$\vec{a}_t$	-	m / sec^2
aerodynamic power	P	-	J/sec
power coefficient	C_p	-	-
Rotor Inertia	J_r	3.5444e7	$kg \cdot m^2$
Generator Inertia	J_g	534.116e2	$kg \cdot m^2$
Driveshaft stiffness on rotor side	k_r	8.676e8	Nm/rad
Driveshaft damping on rotor side	b_r	6.215e6	$Nm \cdot s/rad$
Gear ratio	N_{gr}	97	-