TECHNISCHE UNIVERSITÄT DRESDEN

FACULTY OF COMPUTER SCIENCE
INSTITUTE OF SOFTWARE AND MULTIMEDIA TECHNOLOGY
CHAIR OF COMPUTER GRAPHICS AND VISUALIZATION
PROF. DR. STEFAN GUMHOLD

Großer Beleg

Reinforcement-Learning Windturbine Controller

Nico Westerbeck
(Mat.-No.: 3951488)

Tutor: Dr. Dmitrij Schlesinger, PhD. Matthew Lennie

Dresden, December 9, 2019

Aufgabenstellung

The following problem formulation is an exact copy from what M. Lennie drafted at the beginning of the work and the only part of this extract which is not formuluated by N. Westerbeck.

Problem

Here at the HFI Experimental fluid mechanics group, we have developed an open source project called QBlade. QBlade is a simulation tool used for testing wind turbines in the hostile environment that they normally operate. We normally tackle problems of aerodynamic or structural optimization but we have also a research focus on the development of the control systems of the wind turbines. We currently have a research effort looking at developing cluster-based controllers building on the work of Professor Bernd Noack who is a guest professor at our group. In the last year or so (Nair, A. G., Yeh, C.- A., Kaiser, E., Noack, B. R., Brunton, S. L., & Taira, K. (2018). Cluster-based feedback control of turbulent post-stall separated flows. Journal of Physics Fluid Dynamics, (M), 1-32. Retrieved from http://arxiv.org/abs/1809.07220). AI projects such as openAI have enabled he rapid development of neural network in the field of control using reinforcement learning. The goal of this project is to use QBlade as a wind turbine simulator and attempt to control the pitch and rotor speed in a way that doesn't cause the wind turbine to shatter but instead to yield energy, i.e. reward and death condition. This first stage of work should be considered as exploratory but will hopefully open up avenues of controlling active flow control elements such as flaps.

Tasks

The major tasks of the project are as follows:

- Build up and interface between QBlade and python the model code so that an external code can run as a controller within a QBlade simulation.
- Gain a rough understanding of the mechanics of wind turbines and their controllers.
- Research reinforcement learning methods suitable for use as a windturbine controller and perform a literature review on these approaches.
- Create a reinforcement learning agent which uses the Qblade interface for controllers to control a windturbine.
  - Inputs to the agent could be defined by the standartized controller input format to Nordex turbines, which consists of 39 real-valued sensor-inputs. However, initial tests can be conducted with whichever inputs are easiest to tackle. If required, further hidden state from the simulation can be exported to enrich data quality. If aiming for industrial quality, more inputs and also sensor faults could be optionally incorporated.
  - Outputs are in a minimum version pitch angles for the 3 blades and turbine torque. Optionally the agent should be able to control active element such as flaps on the blades.
- Optimize the agent to deliver maximum energy yield.
- Optimize under respect of certain boundary conditions (maximum pitch acceleration, maximum power, maximum blade load, blade touching the tower) and optionally other boundary conditions like long term turbine wear.
- If necessary for the training process, scale the simulation to run at a larger scale.
- Implement and attempt to get the agent to perform something close to sensible control of the wind turbine. Optionally evaluate the results against existing controllers and try to outperform them.
- Optionally, create a conference paper, poster or blog post etc.. on the results.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die von mir am heutigen Tag dem Prüfungsausschuss der Fakultät Informatik eingereichte Arbeit zum Thema:

Reinforcement-Learning Windturbine Controller

vollkommen selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Dresden, den December 9, 2019

Nico Westerbeck

Contents

1 Abstract

Recent advancements in Reinforcement Learning (RL) have managed to tackle more and more complex problems, like StarCraft or Go. The range of topics, RL is applicable to, increases. However, so far, these advancements were only partially applied to the area of windturbine pitch and torque control. Current state of the art linear controllers are performing well in maintaining turbine control, and the optimization margin for these is small, as they achieve near theoretical maximum energy output. However, these simple controllers can only deal with a limited amount of data inputs and are struggling to prevent long-term turbine damages. We are evaluating the use of DDPG as an actor-critic reinforcement learning algorithm to solve continuous control of a windturbine, hoping to lay a foundational work for later improvements. We find that utilizing the unmodified algorithm does not solve our windturbine problem, thus propose to extend DDPG with a range of known and novel methods, including prioritized experience replay, death handling and normalization. Finally, we demonstrate that windturbine control is possible, however we do not achieve results outperforming industry controllers.

2 Introduction

With the issue of global warming, we are facing maybe the biggest challenge to our generation. Though weather models have predicted the effect since the 1980s, only in the recent years this topic has been brought into societal focus. CO2 is seen as one of the main cause for a tendency in the global weather to heat up. Since our industrial age, humankind has been releasing CO2 into the atmosphere in large scales. If however we continue doing so the way we are doing now, we will likely increase global temperatures by roughly 3 degrees which will make living on this planet difficult. With the Paris agreement, the United Nations decided to limit global warming to 1.5 degrees, for which massive reductions in CO2 emissions are necessary. As energy production holds one of the biggest parts in the global CO2 emissions, we need to replace CO2 emitting power plants. For this, renewable energies have played a major role, and as a big part of this also windpower.

Winds are a direct result of sunshine. The sun heats different parts of the earth to different temperatures, as the surface of the earth isn't uniform. The air over hotter regions such as close to the equator tends to rise, while over oceans and the poles it tends to fall. Combine this with a vast and complex interplay of humidities, Coriolis effects, mountains, cloud formation and enough co-dependence to fill an entire branch of science and you will get wind. Windturbines operate by drawing energy from this weather phenomenon. Large rotors placed in windy regions which are connected to a generator yield electrical power. As weather is a complex phenomenon, wind tends to come from different directions with different speeds. This makes drawing energy from it more complex, or less efficient. If we were to draw energy from both storms and light breezes with the same turbine, we would not be efficient as the turbine can only draw so much energy that it survives the storm, which projected on the light breeze is very little energy yield. Thus, modern windturbines are built to be adjustable. With these adjustments, a turbine can operate sensibly over a broader range of wind conditions. However, somehow, these adjustments need to be automatically adapted to the inflowing wind. This is the task of the turbine controller. We are attempting to learn such a controller.

Machine learning has developed immensely over the last years. Benefiting from Moore's law, huge amounts of computational power can be used to train more and more complex models. Machine learning is generally split into three big branches. Supervised learning, unsupervised learning and reinforcement learning. In supervised learning, a function approximator is fitted to model properties of a dataset. In unsupervised learning, properties in a dataset are measured, compared or grouped. In reinforcement learning, the dataset is an environment and it is explored interactively. We are trying to learn a controller for a windturbine, so in theory we could use both supervised an reinforcement learning. If we were to use supervised learning, we would need some precomputed data, which could only be generated by already running a windturbine with a controller and measuring the data. As we however want to learn this controller, it is not present at the start of supervised learning, and imitating known controllers could at best get exactly as good as the known controller. So we are using reinforcement learning.

With the combination of both, we aim to achieve insights into the applicability of reinforcement learning to wind turbine control tasks. We are aiming at laying a foundational work, not necessarily yet outperforming state of the art control systems. Our hope is that the rising performance and more sophisticated algorithms are able to solve pending problems in wind energy now or in the near future.

3 Background

This section aims at facilitating the background knowledge necessary for understanding this work. We will first introduce windturbines and common terms around this area of research. Then we will motivate our work precisely. We will give insights on the simulation tool we used and then derive the reinforcement learning algorithm we decided to use. Lastly we will present another paper which tried to combine reinforcement learning and windturbines.

3.1 Windturbine control

There are two major types of wind turbine designs, Horizontal Axis Wind Turbine (HAWT) and Vertical Axis Wind Turbine (VAWT), which differ by their rotation axis. In this work, we will only look at HAWT as these are more commonly used, in our work all mentions of windturbines mean HAWT type turbines. Such a turbine is made up by 3 big components, a tower on which a nacelle is rested which itself has a rotor in front. The joint between tower and nacelle allows for rotation to turn the rotor into the wind, and the angle between inflowing wind and nacelle orientation is called yaw angle. Except for recent ideas [HLD], the best value for the rotational direction was to always yaw the rotor directly into the wind, and for the sake of simplicity we will omit this control parameter from our simulation and leave it at 0 degrees with wind facing straight onto the rotor. In the nacelle there is, most importantly, an electrical generator, which can be given a specific torque. Torque is a word for the amount of rotational force, here in opposite direction to rotation. This force will slow down the rotor and at the same time generate energy. If generator torque is approximately equal to the torque created by wind, the two cancel out and leave the rotor at constant rotational speed, which is what we want to achieve. This torque is one of the two more important control parameters, together with blade pitch. The blades are designed to operate on maximal aerodynamic efficiency when they are pitched close to 0 degrees, increasing the pitch angle will (except for possible slight improvements in the first few degrees) reduce aerodynamic efficiency of the rotor. The main use case of blade pitch is to protect the turbine from damage, as high wind speeds can generate higher loads than the turbine could handle. Our generator has a maximum torque limit, and if the inflowing wind generates more than that torque, the turbine would spool up faster than what it can handle. We want to pitch the blades as little as possible so we can maintain maximum generator torque.

As described in [BJSB, sec 8.3], the normal controller design for windturbines divides the operation of a turbine into two parts, below rated and above rated. To explain this, we added reference plot of the windturbine model which we will use in our further testing, the NREL 5MW reference turbine [JBMS]. Figure 3.1 describes several properties of that turbine in relation to windspeed. Rated windspeed is a windspeed, in which the turbine gives maximum energy on highest aerodynamic effiency. In our example figure, this point is at 11.4m/s inflowing wind. At this point, the generator runs on its maximum torque (green graph) and the blades are pitched to the optimum angle of 0 degrees (red graph). At this point, the tip-speed-ratio (purple graph) is at its designed spot. When designing the turbine blades, this tip speed ratio plays an important role. Below that windspeed, the rotor generates less rotational torque than the maximum generator torque, thus the generator torque needs to be reduced to not slow the rotor down too much. Slowing the rotor down complicates the energy generation process, and though variable-speed wind turbines allow for some play, all windturbines are limited to a certain operational range. Fixed-speed wind turbines can only run on one single rotational speed and thus need to change the torque more aggressively. In our reference turbine, you can see how the rotor speed is kept above 6.9 rpm, which is the cut-in speed of the NREL 5MW turbine. As the generated power of a windturbine is proportional to

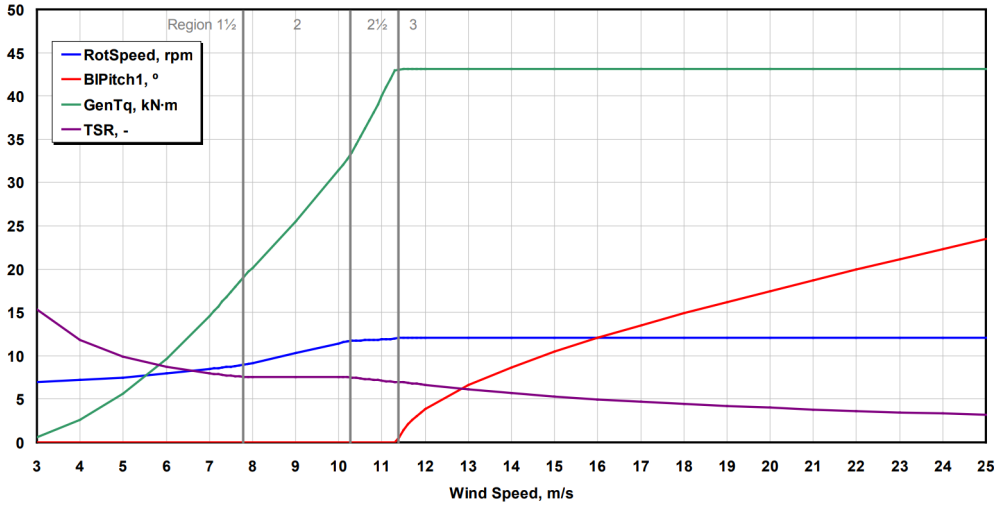


Figure 3.1: Control curve of the NREL 5MW turbine, from [JBMS]

its rotor speed times generator torque, we produce less than our rated energy in this area. Above rated windspeed, the blades need to be pitched out to stop the rotor from spinning at faster speeds than what the turbine was designed for. See the red line in the reference plot - blades are kept at zero below rated and from then on control is done only through blade pitch change. The generator torque can be kept at maximum in this area and such will also the energy production stay at maximum. To protect the turbine from damage, aerodynamic efficiency of the rotor is reduced by turning the blades along their long axis and thus moving their angle of attack against the incoming air into a less optimal range. This is done in a way that the rotational speed of the rotor stays exactly constant.

Additionally to this trivial control, optimizations to reduce vibrations and stress on the structures are implemented. The blades can be pitched individually and quickly enough to allow for different blade pitches during a single rotation of the rotor, which can be used to reduce the turbulence that hits the tower or to account for different windspeeds closer to the ground and further up in the air. Also, both generator torque and blade pitch can be used to counteract natural resonant frequencies of the structure, reducing material stress through extensive swinging of structural parts. In our 5MW reference turbine, only blade pitch is used to counteract resonances.

Usually this control is implemented by two PID controllers, which are hand-designed as described in [BJSB, sec 8.4]. PID-Controllers in general have one real-valued input over time and deliver an output. This is done by calculating an error term between the measured input and a desired input. The output is calculated as a mixture of the error itself (P - proportional to the error), integrating the error over the last few time-steps (I - integral of the error) and calculating a derivative to the last time-step (D - derivative). The output is the sum of these 3 terms each factorized with a constant factor. There are well proven formulas how to choose these parameters and also some adaptive methods, but control theory usually chooses these parameters by hand. However back to windturbines. Usually, we implement two PID controllers to control a turbine, a torque and a pitch controller. Below rated windspeed, the controller for torque is active, above rated the one for pitch. The parameters to those controllers can be calculated according to the laws of control theory. Sometimes, this strong split into two operational ranges is relaxed a bit, as in the 5MW controller, the pitch part counteracts resonances also below rated speed. From this simple concept, the resulting controllers work reasonably close to theoretical maximum already. Our reference turbine has a peak power coefficient of 0.482 - meaning that at optimal wind speed, 48% of the energy of the wind is converted into electricity. In fact, the theoretical maximum is given by the Betz limit of 59.3% and if we account for unavoidable electrical and frictional losses, the realistic maximum is even lower. So there is not much potential to be alleviated. Also, only a fraction of this potential lies in the responsibility of the control circuit. Blade design has so far been the major area for improvements.

Still, in this work, we are trying to replace these hand-designed controllers with a reinforcement-learning algorithm. We are not expecting to exceed the performance of a industry controller. So why do we even do this?

3.1.1 Motivation

As you might have seen before, all controllers follow a certain style. They take a single input variable, such as rotor speed, and adjust a single output such as torque. This type of controller is called Single Input Single Output (SISO). SISO controllers are easy to design and work reliably. They can also be interconnected, such as with the pitch controller which gets a vibration input and a rotor speed input to combine them to one output. The main challenge with wind turbine control at the moment is however not to optimize for power, but to optimize for life-time [vKPN$^+$, Chapter 4]. The controllers are dealing well with adjusting energy to a good level, but they are missing out on keeping the windturbine intact over a long life time.

Also, the outputs get more complex. Back in the early days of wind turbines, you could control 3 variables - pitch, torque and yaw. Nowadays, we can control each of the blades individually and with a quick enough response time to wiggle the blades around within one single rotor rotation. Research on adding active control elements to the blades such as flaps is happening, which would add many more control parameters. You might have seen flaps on an aircraft wing, extending the length of the wing or adding gaps at takeoff and landing to account for the low wind speeds hitting the aircraft wing in these flight scenarios. This could also benefit wind turbines either for higher efficiency on low wind speeds or to furthermore reduce blade loads. However building a good controller for them is challenging due to the high number of interplaying parameters. Tackling all this with SISO is possible, but a lot of manual work.

Next, the inputs also get more complex. It is getting cheaper to install high-precision sensors, and many of them. Wind measurement techniques such as LIDAR based anemometers are able to efficiently and precisely predict incoming wind situations. Load measurements can be done in real-time across a turbine blade and in theory, it would be possible to incorporate these into control systems. However, accommodating all these input parameters in SISO systems is difficult as we would have to hand-model each of the inputs and how exactly it ends up in the output.

We hope to lay a foundational step in solving these issues by using an end-to-end reinforcement learning algorithm which is capable of learning a wind turbine control scenario with many inputs and outputs. Also, we must admit that we are generally interested in playing around with reinforcement learning, it's fucking cool.

3.2 QBlade

Our data source in this work is the open-source simulation tool QBlade [MWP$^+$] developed at TU Berlin. QBlade is an accessible and performant tool with the primary purpose of designing and simulating wind turbines in a graphical user interface. Its simulation results are on par with current state-of-the-art proprietary simulation tools and it yields good computational efficiency. It uses an algorithm which approximates the 3D blade structures with 2D structures and corrects the results through several error terms. Though a full 3D CFD simulation would yield higher accuracy, the increase in computational power needed doesn't justify the accuracy improvement yet. To be used in reinforcement learning, we designed an interface, over which the QBlade simulation can be embedded into an external environment. Most machine-learning frameworks are written in python, so we decided to compile QBlade into library format and expose the most fundamental functions to allow it to communicate with any programming language that can load libraries. During this work, David Marten of TU Berlin was of great help, implementing

all necessary changes to QBlade so it can be compiled as a library, thank you for this! As python has a ctypes interface to load c-code, we could link a python agent to the QBlade C++ environment.

QBlade allows different simulation scenarios, for our testing we decided to only use the NREL 5MW [JBMS] turbine with the default structural model and using all implemented correction mechanisms to achieve the most realistic data possible. As reference data to this turbine is easily available and publicly published by NREL, this allows for good cross-validation.

The interface is made of 7 functions, which allow for loading a project, resetting the simulation, setting controller inputs and getting environment measurements and advancing the simulation a time-step. The observations returned in getControlVars match those that are visible to standart industry controllers, which in turn are modeled after what can be measured on a real windturbine. We have 23 observations and 5 actions.

After a while of experimenting with the unrestricted simulation, we added some bounds to our control parameters and hid some observations and actions. Restrictions to our final version are described in section 5.1.

3.3 Reinforcement learning

So, why are we using reinforcement learning? We have posed the challenge earlier, that we want to learn a control system with a high number of inputs and outputs. Reinforcement learning has proven to be able to tackle high-complexity problems, both in input and output dimension. There have been successes in learning to play atari games based on pixel inputs [MKS$^+$a] and outperform humans, there have been successes in playing games with very long-term strategies such as go [SHM$^+$], and research is continuing to improve. Because of all these successes, we try to automatically learn a windturbine controller.

Unfortunately, reinforcement learning isn't possible without a good amount of maths, so the following sections will be mainly mathematical.

3.3.1 Environment assumptions

At first, we need to do some assumptions on our environment ε. The concept environment encapsulates the reality or simulation that the reinforcement learning algorithm works in. This could be an actual cyber-physical system or a simulation. The first assumption to this environment is that this system operates over time and we can discretize time into time-steps. Our environment here is the simulated wind turbine, but it could be a computer game [LHP$^+$], a physics simulation[BCP$^+$] or even an actual existing wind turbine [KJT]. In every time-step, the environment supplies us with an observation x and a scalar reward r for taking an action a in that state. In theory, the full trajectory $(x_0, a_0, ..., x_t)$ could be needed to describe the full state of the simulation at time-step t s_t. However, the algorithms we looked at assume that the state progressions can be modeled as a Markov Decision Process (MDP) of observations, as in what will happen at time-step t+1 is only dependent on the observation at time-step t, not also on for example t-10. In other words, in MDP, the state transition probability $p(x_{t+1}|x_t, a_t)$ is fully descriptive, as in it is equal to $p(x_{t+1}|x_0, a_0, ..., x_t, a_t)$. Because we assume the observation x to be fully descriptive, we will use the term state as equivalent to observation $s_t = x_t$. In reality, the observed state doesn't have to represent the entirety of the state of the simulation, in fact it usually is only a small subset. For our wind turbine example, the observed state is limited to what can be measured with sensors on a wind turbine, in an arcade game it could be the pixel output or in a physical simulation it could be some key values in the simulation. We denote the probability of a state transition to a specific s' as $p(s'|s, a; \varepsilon)$ and the distribution over all states s' as $P_{s' \sim \varepsilon|s,a}$ and when s, a is clear, we omit it.

The reward $r(s, a)$ is a scalar value which judges the action a taken in the current state s of the simulation.

A higher reward means the action is better than that of a lower reward. Sometimes, the reward function is obvious from the system, as in an arcade game it would surely be the score or in chess it could be 1 for win, -1 for lose and 0 for not decided yet. In both cases, sometimes a certain timespan passes before a reward for an action kicks in, which is also to be learned. In our paper, we will dedicate a section on how we designed our reward function, whereas in general reinforcement learning theory, it is assumed to exist and be supplied by the environment.

Additionally to the state and reward, an environment can optionally send out a done signal d, which indicates a state in which further simulation is not possible and the environment wants to be reset to an initial state. In a chess game, this would be a loss or a win, and with a windturbine that could for example be fatal structural damage.

In every time-step, the agent supplies an action a. This action could be button presses in an arcade game or blade pitch in a windturbine simulation. The mission of the agent is to pick actions which maximize cumulative reward. In other words, it should steer the environment to achieve best performance. Let's write down this mission. We want to maximize:

$$J = \mathbb{E}[r] \tag{3.1}$$

This formula is still a bit incomplete - we didn't yet specify our expectation. And wasn't reward defined on states and actions? But that is because we are still missing some definitions. Luckily, there are some ahead.

3.3.2 Definitions

Additionally to the terms defined by the environment, we introduce some terms which are common to reinforcement learning

- **Policy** A function that maps from states to actions. There are deterministic policies $\mu : S \to A$ and stochastic policies $\pi : S \to P(A)$. If we approximate a policy by a function approximator parametrized with θ, we write $\pi(s;\theta)$ or omit the parameters when clear. We sometimes call a learnt policy *actor*
- **State-density** We define a function similar to [SMSM] which describes how likely it is to be in state s when following a policy π forever. $d^\pi(s) = \lim_{t->\infty} p(s_t = s|s_0, \pi)$. If we for example have a policy which plays Go as well as the author of this paper, states with a lot of enemy stones and very few own stones might happen more often than others, and a state without any enemy stones would be very unlikely be seen at all. Also according to [SMSM] this distribution is independent of the starting state s_0 and only dependent on the policy and of course the environment. In later papers, commonly a discounted version of this is used. $\rho^\pi(s)$ weights future probabilities less and thus can be described as the probability of being in state s *soon* when following policy π. We will omit a precise definition here, you can look it up at [SLH$^+$, Section 2.1].
- **Value** A value with respect to a policy and a starting state s gives the accumulative reward of following that policy from state s until infinitely in the future. $V^\pi(s_t) = \mathbb{E}_{s_i \sim \rho^\pi, a_i \sim \pi, i>t} r(s_i, a_i) = \int_{s_i, i>t} \rho^\pi(s_i) \int_{a_i, i>t} \pi(a_i) r(s_i, a_i)$. As we have a stochastic environment and policy, we have to add an expectation term over actions taken and states resulting from those actions. It is the expected accumulative reward of following that policy. Furthermore a value can be discounted or undiscounted. Discounting means we multiply a discounting factor $\gamma \in [0, 1]$ to rewards in the fashion of $r_0\gamma^0 + r_1\gamma^1 + r_2\gamma^2$..., which results in distant future rewards being weighted less than closer ones.
- **Return** If we measured a value function, that would be the return. Obviously we can't measure infinitely, so a return is only defined on a finite trajectory of state-transitions that we can actually observe, but otherwise is equal to a value.

- **Optimal policy** The theoretical construct of a policy that always takes the best action with respect to accumulative reward is called optimal policy. Though we rarely know this policy, we usually want to know it. In a finite MDP, there must always be at least one policy which gives this highest value of all possible policies.
- **Q-Value** The reward of taking action a in state s plus the value following a policy after that $Q^\pi(s,a) = r(s,a) + \mathbb{E}_{s' \sim \varepsilon} V^\pi(s')$ is called Q-Value. When following the optimal policy, we simply write $Q^*(s,a)$. Usually, our Q functions are discounted similar to how we discount values. Note that we can define Values and Q-Values on each other: $V^\pi(s) = \mathbb{E}_{a \sim \pi} Q(s,a)$. If we approximate Q by a function approximator parametrized with θ, we write $Q(s,a;\theta)$ or omit it in case it is clear. We sometimes call a learnt value estimation *critic*.
- **Actor-Critic** Actor is just another name for a learnt policy and critic just another name for an estimated Q-Value or Value function. These two frequently appear together, if that happens we are using an actor-critic type algorithm.

Now, with all these definitions at hand, we can specify our mission a bit better. We want to find a policy $\pi = \mathrm{argmax}_\pi J$ which maximizes expected value

$$J^\pi = \mathbb{E}_{s \sim \rho^\pi, a \sim \pi} r(s,a) = \int_S \rho^\pi(s) \int_A \pi(s) r(s,a) \tag{3.2}$$

In other words, we want a very high probability of being in a state (ρ^π) which might give good rewards, and in that state we want a high probability of taking an action (π) which might give a good reward. In our wind turbine, rewards could for example be the all the costs and incomes summed together, so at every point in time we could count income by energy yield or costs because of maintenance (with not so optimal policies costs for rebuilding the entire turbine). This work would be easy if we could directly calculate this equation but clearly, there is a bit of a problem ahead. Firstly, we don't know ρ as we don't exactly know the environment we are modeling. We can observe some rewards, states and actions, but only very few of them. We might then see that using 0.3MN of torque at rotational speed 0.8 rad/s yields good energy yield, but how about rotational speed 0.9 rad/s? We might have never seen that. Also, rotating the turbine at 2 rad/s might be a good strategy to immediately achieve high power outputs, but we might need to fix our turbine after running this policy for just a few days and incur a heavy penalty. So we don't know how rewards will spread out in the future. Luckily we have defined Q-Values above. We could use them somehow, so let's discuss how to learn them. If you already know about Q-Learning, you can skip the next section.

3.3.3 Q-Learning

At first we will present a very basic algorithm for a Q learning agent, which is based on the basic Bellman-Equation [Bel]. It is proven, that

$$Q^*(s,a) = \mathbb{E}_{s' \sim \varepsilon | s,a}[r(s,a) + \gamma \max_{a'} Q^*(s',a')] \tag{3.3}$$

converges to the optimal solution when using iterative value updates. We will quickly explain this Bellman Equation. This equation uses the notion of a state-value function, also called Q function. In the Bellman Equation, it describes the total accumulated return until infinity taking an action a in state s, receiving reward r, ending up in state s' and then following the optimal policy afterwards. γ acts as a discounting factor just as in our return function. It is even possible to implement an algorithm from this. We could create a lookup table for all s and a, initialize it all zero and whenever observing a state transition, we could update the table at that state and action after the above notation. This iterative Q update is proven to converge to the optimal solution.

With the help of this function, we can compare different actions in our current state. However, this table is only computationally feasible when we have small finite state and action spaces. If any of the two spaces is big or even continuous, we have to replace the table by an estimator for Q.

Even with an estimated Q, though no longer guaranteed, we can still converge towards the optimal solution. For training an estimator, we need a loss which we can derive. For this, we define our learning target for learning step i as

$$y_i = \mathbb{E}_{s' \sim \varepsilon}[r + \gamma \max_{a'} Q^*(s', a'; \theta_i)] \tag{3.4}$$

For the moment, let's assume we can still calculate $\max_{a'}$. In fact, in a small finite action space we could just try all possible values of a. Our y is then the current reward added to the highest possible rewards in the future, with other words exactly what we want to have as Q. We use the parameters θ_i in the target, so in theory, when deriving a later loss, we would also have to derive over our targets. However, in Q learning, this is commonly ignored.

Now, on our learning target y, we can define a loss as

$$L(\theta_i) = \mathbb{E}_{s \sim \varepsilon}[(y_i - Q^*(s, a; \theta_i))^2] \tag{3.5}$$

This already looks close to some squared error term. To calculate the expectation value over the environment ε, we need to do some form of Monte-Carlo experiment, in which we approximate that expectation value by repeatedly drawing samples from the underlying distribution. Luckily, we can easily draw samples from our environment, and we can even reuse old samples. Thus we store all samples in a replay buffer and uniformly draw from it to resemble a Monte-Carlo experiment.

Deriving for θ_i and applying the chain rule, we get the loss derivative

$$\Delta_{\theta_i} L(\theta_i) = \mathbb{E}[(y_i - Q^*(s, a; \theta_i)) \Delta_{\theta_i} Q^*(s, a; \theta_i)] \tag{3.6}$$

Using our Monte-Carlo like batch from the replay buffer, we end up with stochastic gradient descent.

Thus we just derived a method to train a Q function under the assumption of an optimal policy. In [MKS+a], the authors worked on a finite action space and could use the greedy policy $argmax_a Q(s', a)$, coming reasonably close to an optimal policy. They call their algorithm Deep Q-Networks (DQN), and with some improvements which we omitted here, they are able to play atari games above human level. We are presented with an infinite action space though, which is why we can't implement a greedy policy without high computational expense. So, for our work, we have to also learn a policy and not just a Q estimator. So let's try to derive a way to learn a policy.

3.3.4 Policy Gradients

In this section, we will at first give an intuitive explanation of a basic algorithm which forms the basis for many modern reinforcement learning algorithms, including the one we chose. The algorithm is called REward Increment equals Nonnegative Factor x Offset Reinforcement x Characteristic Eligibility (REINFORCE) (yes that is the name) and was originally presented by [Wil]. Then we will add a Q term to it according to [SMSM] and call it Policy Gradients (PG). [DWS] generalized that term and made an off-policy version of policy gradients, which we will not discuss in detail. That version was then used by [SLH+] and the stochastic policy was replaced by a deterministic one, the algorithm was called Deterministic Policy Gradients (DPG) and we explain it in section 3.3.5. This algorithm finally lays the basis for Deep Deterministic Policy Gradients (DDPG) by [LHP+], which is what we used in our work and which we explain in 3.3.6.

Let's assume we have a probabilistic policy $\pi : S \to P(A)$ which is parametrized by θ. An intuitive way of improving this policy could be to increase the gradient proportional to the reward an action yielded:

$$\theta_{i+1} = \theta_i + \alpha r(s, a) \Delta \pi_{\theta_i}(a|s) \tag{3.7}$$

We could do this every time we see a state transition and we would end up with an iterative way of improving our policy. In fact, this is the basic idea of the REINFORCE algorithm, with some additions. However, we might not know rewards for any action, state or have noisy rewards. We derived Q-Learning in the section before and it would be nice to use the advantage of being able to look into the future for our policies. We can simply replace r with Q, thus have an estimation which action will give which return in this step. So, with our newly gained Q, for all actions and states, do

$$\theta_{i+1} = \theta_i + \alpha Q^*(s,a)\Delta\pi_{\theta_i}(a|s) \tag{3.8}$$

This way, actions with higher Q values will receive a higher gradient step. As Q is constant with respect to θ we don't have to derive along Q here. Problematic is though, that the Q term we used, always took the best possible actions, which could be vastly different from the actions under our policy. So for the optimal policy, jumping down a bridge is the best possible strategy because it also knows how to swim, but we might not know how to swim yet and should better not jump down a bridge. Thus, we need to reformulate Q to account for our policy instead of the optimal policy

$$Q^\pi(s,a) = \mathbb{E}_{s'\sim\varepsilon,a'\sim\pi|s,a}[r + \gamma Q^\pi(s',a')] \tag{3.9}$$

Now, in theory, we can not regard Q as constant with respect to θ anymore, because it is dependent on the policy. [SMSM] claim though that it can be omitted and still regarded as constant. However there is still a problem if we integrate this new Q into our policy updates. We assume that we use the same π for exploration while doing training, so this policy is responsible for taking actions in the environment and also responsible for which experiences we see during our training. So let's assume the policy is poorly initialized at the beginning of our training and gives action jump a probability many times as high as action walk on, though taking action walk would yield a better return in our (already perfectly trained) Q function. Now, as we are taking action jump many times more often than action walk on, we will also perform gradient updates on action jump more often in our Monte-Carlo draw. Though with each update step, we only get a tiny value in comparison to the value of the better action, the higher frequency of updating that action will compensate for it and summing all the tiny values together leads to jumping getting an advantage over walking in this update strategy. To correct for this oversampling bias, we could employ a trick which is called importance sampling. Because we know the bias from the probability distribution of our policy, we divide that probability from the update term:

$$\theta_{i+1} = \theta_i + \alpha Q^\pi(s,a)\frac{\Delta\pi_{\theta_i}(a|s)}{\pi_{\theta_i}(a|s)} \tag{3.10}$$

We might remember from calculus that $\frac{\Delta x}{x} = \Delta \log x$ so we simplify to

$$\theta_{i+1} = \theta_i + \alpha Q^\pi(s,a)\Delta \log \pi_{\theta_i}(a|s) \tag{3.11}$$

So, let's sum up what we have done before and formulate our goal again. Before, we ignored for brevity that we are dealing with expectation calculations and that we are doing these updates under a stochastic environment, so to write it down formally correct, we need to remember the density function of a state under a policy from section 3.3.2. Using the discounted version approximates the Monte-Carlo experiments we are doing better than the undiscounted version, as effectively we do not train forever on one constant policy, but change it frequently, so we will use ρ^π to approximate how likely we are in a state s when using policy π.

Having this ρ, let's write our objective function J which we want to maximize.

$$J(\pi) = \int_S \rho^\pi(s)\int_A \pi(s,a)Q^\pi(s,a)dads = \mathbb{E}_{s\sim\rho^\pi,a\sim\pi}[Q^\pi(s,a)] \tag{3.12}$$

This looks similar to equation 3.2, just that we use Q now.

[SMSM] proves that deriving this yields

$$\Delta J(\pi) = \mathbb{E}_{s\sim\rho^{\pi},a\sim\pi}[\Delta \log \pi(a|s) Q^{\pi}(s,a)] \tag{3.13}$$

This equation is very famous in reinforcement learning and is called *policy gradient*. And by coincidence, it is equivalent to what we intuitively derived in equation 3.11. We decided to skip the actual proof and use this intuitive explanation instead, and we used the intuitive explanation in [SB, Chapter 13, Section 3, after Equation 13.8] to explain the nature of this policy gradient equation.

In reinforcement-learning, it has proven helpful to do off-policy learning, in which another policy β than the one being trained can generate experiences. We would then reformulate our expectation to $\mathbb{E}_{s\sim\rho^{\beta},a\sim\beta}$. Thus, we could store past experiences in a replay buffer and learn from that. In fact, we were already able to use a replay buffer in the Q learning part. If we wanted to do that on our policy however, we would need to correct for the bias of the other policy as well with importance sampling. However we will not derive stochastic off-policy updates, because using the DPG algorithm, this problem is solved. If you want to learn about Off-Policy Actor Critic (Off-PAC) with stochastic policies, we recommend to read [DWS]

Additionally to this problem, another problem will be solved by DPG as well: If we remember our Q expectation term, in which actions were sampled from the policy $\mathbb{E}_{a\sim\pi}$. We needed this expectation term, because we were dealing with stochastic policies. If instead learning deterministic policies, we can exclude the policy expectation from the equation. But before going through all the advantages of deterministic policies, let's first explain it.

3.3.5 Deterministic Policy Gradients

Before [SLH$^+$], stochastic policies were preferred because the stochastic nature aided exploration and because there was simply no theory on how to derive deterministic policy gradients. The idea behind deterministic policy gradients is to replace the stochastic policy $\pi : S \to P(A)$ with a deterministic one $\mu : S \to A$. Also, we are switching from small finite action spaces to continuous action spaces. Now, updating this policy according to the equations above does not work anymore due to two problems. At first and most significantly, above we needed all a or at least some form of maximization to calculate the full gradient of $\pi(a|s)$. This is difficult with a continuous action space now. Furthermore our importance sampling trick, our correction for overestimation bias based on the bias of the policy, does not work anymore. We don't now the bias of that policy, because it is deterministic and will yield only a single action for any state.

So, following the policy gradient has become difficult. However, let's have a look at our Q function. We can replace our stochastic policy with a deterministic one and get

$$Q^{\mu}(s,a) = \mathbb{E}_{s'\sim\varepsilon|s,a}[r + \gamma Q^{\mu}(s', \mu(s'))] \tag{3.14}$$

Now, our Q is only calculating an expectation on the environment distribution, not anymore both on $s' \sim \varepsilon, a' \sim \pi$. This is an advantage, because computing only one expectation requires less samples and we do not have to worry for any bias from the policy in Q-updates. When we do our Monte-Carlo experiments to update it, we can take our generated samples from the environment and assume they are modeling the underlying environment.

If we are visualizing our Q as a function of state and action, we get a fully differentiable surface. If we derive this surface along the action dimension, we get a gradient which points in the direction of a better action for this state. We could directly follow this gradient and calculate our policy update from it:

$$\theta_{i+1} = \theta_i + \alpha \mathbb{E}_{s\sim\rho^{\mu_{\theta_i}}}[\Delta Q^{\mu_{\theta_i}}(s, \mu_{\theta_i}(s))] \tag{3.15}$$

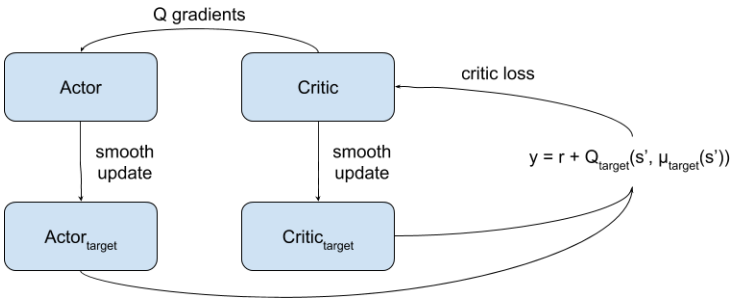


Figure 3.2: DDPG Network

[SLH$^+$, Theorem 1] proves that this gradient always exists. We now managed to remove any expectation that is related to our action space from our gradient updates. This is beautiful because at first, we do not have to sample over many actions to learn this expectation. But more importantly, this term also works off policy, and without importance sampling. We can use a different exploration policy β and replace our expectation term $\mathbb{E}_{s\sim\rho^\mu}$ with $\mathbb{E}_{\rho^\beta}$ here completely unpenalized. Why is this so? [DWS] proved off-policy gradients possible, but had to use importance sampling to correct for the bias in β. In fact, generally, it is not possible to do this replacement. However, [SLH$^+$, Section 4.3] formulated a compatibility theorem. They proved, that a linear function approximator Q^ω which minimizes the MSE between Q^ω and Q^μ will retain the gradient, even when trained on samples generated by β, and called this Q *compatible*. However, in practice we are not explicitly minimizing that loss and neither are we using linear function approximators but usually nonlinear multi-layer perceptrons. [BPS$^+$] proved that using a non-linear function approximator to predict the value function $V(s)$ will converge towards a local optimum. As we can formulate our Q by using V, we transfer this to our scenario. However, DPG is not definitely proven to converge for a non-linear function approximator, as this transfer was, to our knowledge, never fully formulated. However, in practice, it is still trained off-policy. In our own experiments, we found that using a wildly different policy for exploration than for training did in fact result in worse behavior, and [FMP] observed the same behavior. They proved DDPG to not function truly off-policy due to a phenomenon which they call extrapolation error. This phenomenon describes inaccuracies due to poor generalization in the Q function on unseen values. When a target policy predicts such an unseen action, the Q function is likely to overestimate that action. Although there is some evidence against training DDPG off-policy in general, we will still do so because we are not training that far off-policy. We utilize samples from our own, slowly changing policy in the last 100k steps. Also, empirically it delivered good results in the past.

So let's restate our objective function for the actor.

$$J(\mu_\theta) = \int_S \rho^\beta(s) Q^\omega(s, \mu_\theta(s)) ds = \mathbb{E}_{s\sim\rho^\beta}[Q^\omega(s, \mu_\theta(s))] \tag{3.16}$$

Q^ω being a nonlinear function approximator trained under β.

3.3.6 DDPG

DDPG, first presented by [LHP$^+$], is a combination of DPG and DQN. The authors tackle several problems in the two algorithms to combine them. At first, in traditional Q learning, updates to the Q network parameters are based on a target y calculated on the Q network parameters themselves $\text{loss}(Q) = (f(Q) - Q)^2$. This introduces instability and to tackle this, [MKS$^+$a] introduced a second Q network Q', for which the parameters are held constant over a period of updates. They calculate the learning target y based on the second network Q', which is why they call Q' the *target network*. After a number of steps, the parameters from the actual Q network are then copied over to the target. The

stability gain in their algorithm depends greatly on the number of steps after which to perform an update. In DDPG, they instead use smooth target updates, where every step, the parameters of the actual network fade over to the target: $\theta_{target} = \tau\theta + (1-\tau)\theta_{target}$. The fade-over parameter τ was recommended to be set to a small number like 0.01. Additionally to the target Q-network, they also found that a target policy μ' improves stability, and they update it the same way as the main policy. But this is a lot at once. Let's have a look at the Q update functions

$$y = r(s,a) + \gamma Q'(s', \mu'(s'; \theta^{\mu'}); \theta^{Q'}) \tag{3.17}$$

$$L(\theta^Q) = \mathbb{E}_{s\sim\rho^\beta, a\sim\beta, s'\sim\varepsilon}[(y - Q(s,a;\theta^Q))^2] \tag{3.18}$$

We sketched an overview in figure 3.2. Remember that actor was another name for policy and critic another name for Q estimator. Notice how for calculating the Q targets, they used both the target actor μ' for generating actions which are then evaluated by the target critic Q'. This is their version of the improvement over DQN for their instable Q function, as they also observed instabilities. Note how the Q function here does not really predict the value of following a specific policy anymore, as before we could always denote which policy Q was trained on. Also, this Q is approximated by a non-linear function approximator instead of the linear one from compatible DPG. [LHP+] do not provide a proof of convergence for this new, mixed Q update, but instead show empiric proof that it works. So, we have an off-policy way of improving the critic now.

$$J(\theta^\mu) = \mathbb{E}_{s\sim\rho^\beta}[Q(s, \mu(s;\theta^\mu); \theta^Q)] \tag{3.19}$$

Let's derive these two equations. First the policy gradient:

$$\Delta_{\theta^\mu} J(\theta^\mu) = \mathbb{E}_{s\sim\rho^\beta}[\Delta_{\theta^\mu} Q(s, \mu(s;\theta^\mu);\theta^Q)] = \mathbb{E}_{s\sim\rho^\beta}[\Delta_a Q(s,a;\theta^Q|a=\mu(s))\Delta_{\theta^\mu}\mu(s;\theta^\mu)] \tag{3.20}$$

Intuitively, this new policy gradient equation follows the curvature of the Q function directly. Policy gradients point in the direction where the Q function sees the highest values. Our new Q update looks like

$$\Delta_{\theta^Q} L(\theta^Q) = \mathbb{E}_{s\sim\rho^\beta, a\sim\beta, s'\sim\varepsilon}[(y - Q(s,a;\theta^Q))\Delta_{\theta^Q} Q(s,a;\theta^Q)] \tag{3.21}$$

Similar to other reinforcement learning algorithms, we will again not compute the full gradient but use stochastic gradient ascend with batch learning. To rephrase gradient ascend to gradient descend, which is commonly implemented in machine learning frameworks, they multiply the equation with -1. In fact, in the original DDPG algorithm, they used Adam [KB] optimization instead of stochastic gradient descend. As activations, they used relu [Aga]. Their network architecture for the actor consisted of two fully connected layers with 400 and 300 units and instead of relu they used tanh activations in the last layer to bound actions. They had a version which learnt upon pixel outputs and used convolutional layers, while we only look at the low-dimensional version. For the critic, they also used a two layer variant with 400 and 300 units. State inputs were fed through the whole network, while actions were concatenated in onto the second layer, skipping the first. They used batch normalization between the layers and L2 regularization on the weights.

Furthermore, to aid exploration, they added noise in form of a Ornstein-Uhlenbeck process (OU-Noise) [UO]. This noise is a correlated noise which is sampled over time. Instead of the uncorrelated version, this correlated noise can step further away from the noise center (parametrized by μ in their paper). It consists of a mixture of going towards this center (θ) and going into a direction drawn from a normal distribution (σ). As we use the same symbols for other things in our work, we will not adapt their notation.

For readability, we added a copy of the entire DDPG algorithm in Algorithm 1, with minor adjustments to fit our notation.

Algorithm 1: Vanilla DDPG algorithm

Randomly initialize critic network $Q(s, a; \theta^Q)$ and actor $\mu(s; \theta^\mu)$ with weights θ^Q and θ^μ
Initialize target networks Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer $\mathcal{R}$
for *epoch = 1, M* **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for *t = 1, T* **do**
    Select action $a_t = \mu(s_t; \theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in $\mathcal{R}$
    Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from $\mathcal{R}$
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}; \theta^{\mu'}); \theta^{Q'})$
    Update critic by minimizing the loss: $L = \frac{1}{N} \sum_i (y_i - Q(s_i, a_i; \theta^Q))^2$
    Update the actor policy using the sampled policy gradient:
    $\Delta_{\theta^\mu} J \approx \frac{1}{N} \sum_i \Delta_a Q(s_i, a; \theta^Q | a = \mu(s_i)) \Delta_{\theta^\mu} \mu(s_i; \theta^\mu)$
    Update the target networks:
    $\theta^{Q'} \leftarrow \tau\theta^Q + (1 - \tau)\theta^{Q'}$
    $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1 - \tau)\theta^{\mu'}$
  end
end

3.4 RL on windturbines

We are aware of one high-quality paper [KJT] which tried RL on windturbines already, we want to dedicate a section to it. This section is not necessary for understanding the rest of the work, and we have no references to it outside of this section. If you want to understand why we chose a different approach, you can still read this section. The team behind that paper built a miniature model of a windturbine and used a variation of the REINFORCE algorithm to control it.

Their miniature model was built from cheap and low-scale parts, they simulated a wind-tunnel by attaching several fans in front of a wooden tube and at the end placed their turbine. Figure 3.3 illustrates this. Each of the blades of that miniature turbine is individually controllable and through a correction term, they are able to achieve independent pitch control which is able to pitch the motors to a value based on rotor position. Though this turbine rotates many times faster than a normal, full scale wind turbine, the underlying principles are the same (500rpm vs 11rpm). For us most importantly, they used reinforcement learning to control this turbine.

As training input, they used a single observation in contrast to popular reinforcement learning which operates on high-dimensional observation space. This observation was power output of the turbine, and they treated it as reward to the system, not having any state observations. Their action space consisted of electrical load on the generator and blade pitch, which was uniformly applied to all 3 blades. In fact, they designed their policy as to return a gaussian distribution over these two parameters directly without any input: $\pi : \varnothing \to P(\mathbb{R}^2)$. To run the training, they kept the wind speed constant, let the turbine spool up to speed and then let it run with a set of parameters returned by the policy. They ran these parameters for 2 seconds on the turbine, averaged power output over that time into a reward and then ran an iteration of their algorithm to generate new parameters, starting with random parameters. They first tested the

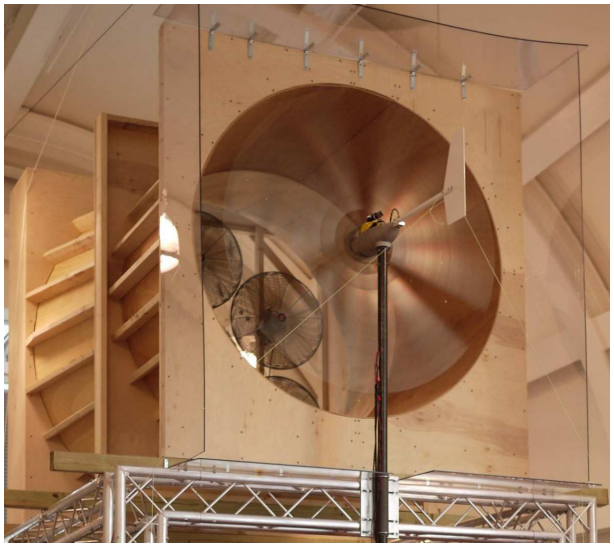


Figure 3.3: The mini wind-turbine from [KJT]

REINFORCE algorithm, which operates without a Q network and trains a policy directly on rewards. This algorithm was discussed above and utilizes the following gradient equation:

$$\Delta_{\pi^\theta} J = \mathbb{E}_{a\sim\pi, r\sim\varepsilon|a} r \Delta \log \pi(a;\theta) \tag{3.22}$$

When reading their work, you will find that they denoted J in their paper as rewards, while we are using r for rewards and J for objective functions. Also, they used ω as a symbol for the action taken, while we use a. Also note how their policy is not dependent on a state but directly predicts actions. They did not explicitly note what they formulated their expectation term on, but we added it here for clarity. Note the similarity to what we discussed in equation 3.7 and 3.13. In addition to this, they came up with an algorithm which they call Trust Region Policy Search, and which is astonishingly different from Trust Region Policy Optimization [SLM$^+$]. With a few additions, they incorporated off-policy learning and second order error approximation into their algorithm, and showed later that it performs nicely. They were able to find optimal pitch and load settings as soon as 15 time-steps after the start of the simulation.

Having described this paper, we want to also summarize what we do differently and why. First, they had a simpler environment. While a full scale windturbine has a strong tendency to destroy itself when outside of their operational margin, their turbine did not fly apart when not applying braking. Also, they were setting electrical resistance, not rotor torque. This system applies a braking force constant to rotational speed times resistance parameter, and thus already acts as a P controller. We were setting torque, so we had to adjust it regularly to rotor speed. In general, their turbine behaved more concavely opposed to our turbine, which with a constant parameter set, would settle to either rotating quickly backwards, forwards or not at all. Because of this, they were able to give the turbine some time at the beginning to stabilize before applying reinforcement learning while we needed to start with an artificially stabilized turbine. Lastly and most importantly, they aimed to find optimal control settings for a constant wind speed. To create a full controller setup from this, they would have to repeat this process for every windspeed and use the resulting look-up table to derive a PID-controller. We however want to create an end-to-end system and want to experiment with the possibly higher expressive nature of predicting actions based on a full state input. We want to create a version as general as possible. Also, as DDPG is a direct successor to REINFORCE it has proven to outperform it in many areas.

Further works on reinforcement learning on windturbine are also [SAZFG$^+$]. This work uses traditional supervised learning to fit a neural network of 10 neurons to the straight line that describes the relation between yaw angle and power output and because what they got looks similar to a Q function they call

it reinforcement learning. [FGFGG] published in a proprietary journal, so we didn't have access to their work.

4 Experimentation

For everyone who is interested in a recap of our design process, in this section we will provide such. We will again present our algorithm in a later section. To understand our design decisions, it might be interesting to read this passage, but you can also skip this passage and just read the rest.

4.1 Gym Experiments

At the start of our experimentation, we took off with a well-known problem and a proven implementation. We used the OpenAI-Pendulum task, in which a pendulum in gravity needs to be held upright by applying a force to it. OpenAI Gyms generally serve as benchmarks for reinforcement learning algorithms and most foundational papers and articles evaluate their solutions based on how well they perform in OpenAI Gyms. We managed to find a solution with our DDPG implementation. We hoped that the implementation might yield results on our windturbine environment as well and exchanged the OpenAI gym for QBlade.

The first runs looked marvelous, perfectly holding rated speed, until we found out that those perfect values were prescribed by a setting in the QBlade project file, and disabling this setting marked the beginning of our actual experimentation phase.

4.2 Starting with QBlade

On our first experimentation steps, we did not yet have experience with how the QBlade simulation behaves with a random or not well designed controller. Our observation was, that the simulation was resting between two different states, which we want to call Forward Maniac Mode (FMM) and Backward Maniac Mode (BMM).

FMM is characterized by a rotational speed of ca 5 times rated speed, at 4.3 rad/s. At this speed, for some reason, the rotor will not speed up anymore. However, severe vibrations and sometimes the case of a disjointed blade happen. Especially in the case of a blade flying off, deflection values skyrocket up to several orders of magnitude higher than values seen in normal operation, and with one or more missing blades, the rotor will quickly stop rotating. This state is usually reached by a too low generator torque while not pitching out blades and sometimes stays stable for several thousand steps until a blade falls off.

BMM is characterized by a negative rotational speed, up to -3.8 rad/s. This can be reached by setting a higher generator torque than the aerodynamic torque induced by wind flowing through the turbine and producing lift on the blades. Also in this state, extreme vibrations and sometimes blades flying off can be observed, and like FMM this state is stable until a blade falls off.

In neither of these modes, the simulation is within realistic bounds and the observations received from the simulation at this part are not credible. After we observed that the controller itself wasn't able to avoid these states on its own, we inhibited reaching these states by outputting a death condition when being in negative rotational areas or high positive. After a death, the simulation gets reset to initial state and the controller has more chances to learn.

4.3 Designing reward functions

As QBlade doesn't provide a built-in reward function, we needed to craft our own reward function based on the observed state. We considered several possible variants. We will first introduce all of them and give a more detailed explanation later.

- **Rated speed** Hold a rated speed. Observe current rotational speed s_{rot} and with given rated speed s^*_{rot} calculate reward $r(s) = -\left|\frac{s_{rot}-s^*_{rot}}{s^*_{rot}}\right|$
- **Rated power** Hold a rated power. Observe current power s_{pow} and with given rated power s^*_{pow} calculate reward $r(s) = -\left|\frac{s_{pow}-s^*_{pow}}{s^*_{pow}}\right|$
- **Maximal power** Generate the maximum power possible. Observe current power s_{pow} and with given normalization constant c calculate $r(s) = s_{pow}c$
- **Penalize current stress** Penalize high blade bending. Observe current bending s_{bend} and with given normalization constant c and another reward function $r(s)$ calculate new reward $r'(s) = r(s) - s_{bend}c$
- **Penalize accumulated stress** Penalize accumulated structural stress. With $x = rainflow(s_0...s_t)$ being the rainflow table and $p(x)$ being a function that maps from rainflow outputs to a scalar penalty, calculate $r'(s_0...s_t) = r(s_t) - p(rainflow(s_0...s_t))$
- **Penalize action gradients** Penalize high action gradients. With a_t being the current action, a_{t-1} the last action, c a penalty constant and $r(s)$ another reward function, calculate: $r'(s,a) = r(s) - c(a_t - a_{t-1})$
- **Penalize death conditions** Penalize when a death condition is reached. With $r(s)$ being another reward function, $d(s) \in 0, 1$ being our indicator for deaths and c a constant penalty, calculate $r'(s) = r(s) - cd(s)$

Rated speed is the easiest of the rewards, as holding a rated speed can be achieved by either pitch or torque or a combination of the two. So both a high pitch and a high torque will stop it from running faster than it should. We used this reward for our first working version.

Rated power is a bit more difficult. For reaching rated power, a certain aerodynamic torque is required, so pitching the blades out completely won't deliver rated power. However especially if setting an artificial rated power lower than what the turbine was built for, there is a certain play in how strongly to use blade pitch and how much needs to be done through torque.

Maximal power is an intuitive reward, but might yield unrealistic results. This rewards high power regardless of the stress that is induced to the turbine, so a good policy might spin the turbine at high speeds beyond any turbine specification and then apply maximum torque. So this reward makes more sense when combined with a stress penalty.

Penalize current stress is the easiest stress penalty, where a penalty proportional to the current bending of the structure can be used. As there are several bending modes on most of the components of the windturbine, a combination or selection of modes and components should be taken before. We decided to limit us to out-of-plane bending of the blade tips. This is a simple variant which leaves out bending in the middle of the blades, bending in the rotational plane, vibration or bending of the tower and torsional stress on the shaft.

Penalize accumulated stress is a more complete, but time dependent variant. Rainflow Counting (RFC) is a commonly used method for calculating structural fatigue in windturbines [BW]. RFC takes in a bending signal over time and returns a table with amplitudes and their frequency, as in how often the structure swung how far. It filters out some higher-frequency swinging in between two lower frequency swings and thus also works for structures that swing in more than one frequency. To use it as a penalty,

we needed to create a function that maps from that table to a scalar penalty. This penalty accumulates past stress, so it increases over time. We suspect this penalty to be hard to learn, as it reacts slowly to changes and a once given penalty will never be lifted again. Alternatively, we could imagine a moving-average like solution, in which rainflow fatigue is calculated on a small trajectory of past states and not on the entirety of the simulation.

Penalize action gradients is an option which is useful in contexts where high action gradients are impossible on the real system. High action gradients mean for example turning a blade 90 degrees in a single timestep, which would rupture a normal blade. Instead of clipping them, penalizing them could include action gradients into the learning process. We only expect sensible results with a stateful agent which knows at least the last action it took or one which has recurrent connections, as otherwise the last action taken is unknown to the agent.

Penalize deaths is simple to implement but yields a not differentiable loss function at deaths. Especially because we are learning a continuous value estimation, it will most likely smooth out along a death and likely underestimate the death penalty and underestimate values close to the death. Alternatively we could imagine penalizing getting close to a death condition continuously.

Most of our experimentation we did with rewarding rated speed and penalize deaths, as we expected this reward function to be the easiest to learn. In a real application, crafting a good reward function will be a bigger challenge, but this is out of the scope of this work.

4.4 Aiding exploration

As in the beginning we observed our algorithm to get stuck in either FMM or BMM, we wanted to improve exploration. The term exploration in reinforcement learning means how well the agent explores the state space of the environment. Exploration is a common problem in reinforcement learning, as algorithms tend to stay in local maxima. Intuitively, we are optimizing a policy to move towards an optimal state, while computing what is the optimal state based on what we observed. Thus, if a certain policy already gets locally good values, it will move towards that locally optimal state and never explore out of it, thus the Q network will not know about the good values beyond that.

Additionally, as we discount future rewards, if the slope between the local optimum and the global optimum is too high and wide, the higher return of the global optimum will not propagate until the local optimum because of the discounting done close to it, and though the Q-network knows the global optimum, the gradients along the Q-function still lead into the local optimum.

We suspected the first problem to happen, as neither of the two modes had good rewards, and going out of these modes continuously improves rewards. To feed a network with new values closer to an optimum, a common practice in reinforcement learning is to add noise.

4.4.1 Action Noise

We first tried a Gaussian noise instead of the recommended OU-Noise [UO] from DDPG: $n \leftarrow \mathcal{N}(\mu, \sigma^2)$. We didn't see any improvements over our old behavior, so we switched to OU-Noise. Additionally to the DDPG paper, we let the sigma parameter decay over time, so that the noise gets less when training has proceeded. We hope to only need this noise in the first stage of the training and then, later, see better results. Our decay is parameterized by a start sigma, end sigma and end step and then held constant at the end sigma. We are unsure whether the decay actually benefited our training, but we kept it in nonetheless as it also didn't seem to negatively impact it. We started it 0.1 above and let it decay to 0.1 below the values from the DDPG paper ($\sigma = 0.2, \theta = 0.15$) within the first 50k training steps.

4.4.2 Random exploration

The TD3-paper [FvHM] utilizes a random exploration phase, in which a completely random policy [MB] initially explores the environment, before the actual agent switches in. We implemented the random walk at first through Gaussian noise and then through an Ornstein-Uhlenbeck process. The later version creates less vibration through less excessive changes in control parameters. As we found that both random policies rarely explore a sensible operational range, we later added an expert PID controller, which implements torque control, and combined that with our Ornstein-Uhlenbeck process. We set the parameters to that controller by experimentation and came up with the following equation, only dependent on rotational speed s_{rot}: $\mathcal{C} : (s_{\text{rot}}) \rightarrow (a_{\text{torque}}, a_{\text{pitch}}), \mathcal{C}(s_{\text{rot}}) = (2 \times 10^7 s_{\text{rot}} - 12 \times 10^6, 128 s_{\text{rot}} - 103)$. These actions were clipped as in section 4.6.3. So, effectively, we designed a P controller, and only using this, we could already achieve relatively stable control. As further work, it could be imagined to use an established controller here, but we did not need this for our purposes. The PID controller was easily able to keep the windturbine in a sensible operation range, and without adding noise it resulted in a smooth and constant operation. We summarize this in algorithm 2

Algorithm 2: Expert exploration

Use replay buffer $\mathcal{R}$ from normal training
for *epoch = 1, E* **do**
  Initialize expert controller $\mathcal{C}$ and random process $\mathcal{N}$ for expert exploration
  Receive initial observation state s_1
  for *t = 1, T* **do**
    Select action $a_t = \text{clip}(\mathcal{C}(s_t) + \mathcal{N}_t)$ Execute action a_t and observe reward r_t, death condition d_t and new state s_{t+1}
    Store transition $(s_t, a_t, r_t, s_{t+1}, d_t)$ in $\mathcal{R}$
    On death condition, reset the environment.
  end
end

We did however not really observe better results with expert exploration enabled. We tried combining it with the algorithm from section 4.6.4, because we were suspecting that the large difference between the actor distribution and the expert distribution lead to problems in learning the respective Q distribution. We argued before that we suspect DDPG not to be able to learn completely off-policy but just a bit off-policy. Together with the pretraining algorithm, we had some slight improvements, but these didn't justify the extra computation time needed for this. Also, though an expert controller which yields somewhat sensible results is easy to build for each turbine, this is a step of human interaction and expert knowledge, which machine learning generally tries to minimize in algorithm design.

4.4.3 Parameter noise

As we still didn't see good results, we tried adding parameter-space noise [PHD+] to the actor function, as the paper promised better exploration. We could not observe any improvements and thus deactivated parameter space noise for the rest of the experimentation.

4.5 Zero Gradients

All our efforts to aid exploration by adding noise did not yield better results, so we had a look at our gradients and found that all our actor gradients are zero. The critic had gradients, but only in the later

levels of the net. The problem of vanishing or exploding gradients is typical for deep learning, but not so common in flatter architectures.

4.5.1 Simplifying the architecture

We tackled this problem by at first reducing the number and size of layers from the example implementation to one fully connected layer of 64 neurons in the critic and one fully connected layer of 32 neurons in the actor. Reducing complexity partially tackled the problem of zero gradients, we could then observe small gradients which however vanished over time. Later, we found that two layers of 64 and 32 in the critic and 32 and 16 neurons in the actor also still yielded good gradients with higher generalization potential.

The authors of DDPG trained their networks up to 2.5 million steps, which we can't afford with our computational power, so a simpler network architecture will hopefully also converge faster. However we loose generalization potential with the smaller network. We will not be able to solve the same complexity of tasks as DDPG with our smaller networks. We estimate windturbine control to be a complex problem though. So effectively, we had to look for other solutions.

4.5.2 Normalization

More effectively, we added data normalization. As in our simulation, where our state reflects measurements from very different parts of the simulation, some values in the state-array were 8 orders of magnitude different from others as they were measured in very different units. As most of the high-magnitude values are vibrations, the net could not create a link between these and the state-action, while low-magnitude but insightful observations like rotational speed were likely not considered due to their small absolute values.

We normalized the data so states, actions and reward usually stayed between 1 and -1. This is a normal technique in most of machine learning, but for some reason, in reinforcement learning this is not so common. At times, rewards were clipped [MKS$^+$a] and observations were sometimes, especially when working on pixel inputs, cropped to be of square dimensions. Only in 2016, [vHGH$^+$] brought up the topic. They propose an adaptive normalization mechanism which integrates into a generic RL algorithm and which can normalize all inputs adaptively. Because of the extra implementation effort however, we decided to implement a more straight-forward variant. At first, we used observed states and rewards from the random exploration phase to calculate the 5%, 95% quartile values. We calculated normalization constants from it to linearly normalize those quartiles to $[-1, 1]$. Plain min/max normalization worked less well. As sometimes during random exploration, no extreme values were observed and later extreme values lay far outside of $[-1, 1]$, we instead normalized on replay data of a previous run. Additionally, we added a constant of 3 after normalization because of the phenomenon described in section 4.8.4, but only in later evaluations. We reference to normalization by norm(s), norm(a) and norm(r)

Alternatively to calculating normalization bounds, it would be possible to set them with human expert knowledge for low-dimensional problems, as human knowledge about the possible state space is usually present. In case when human knowledge about input magnitudes is not present, we recommend the mechanism from [vHGH$^+$], or if you are willing to accept the extra computational effort of running once without normalization, our version.

Normalizing our data drastically helped with our problem of zero gradients and we could observe normal learning. We took the penalty of requiring data from a previous run for our algorithm, but in simulation environments it is possible to acquire this data easily. For a real world environment, it is less easy and our method would not be applicable to these scenarios. We however recommend using any normalization when dealing with different magnitude sensor data.

4.5.3 Last-layer actor activations

In the DDPG architecture, the last layer of the actor is activated with a tanh activation. Like sigmoid, this activation returns close to 1 or -1 for high or low activations. These activations commonly produce vanishing gradient problems when their activations are driven close to this range. A big change in input at these borders results in a very small change in outputs, for example $\tanh(2) = 0.9640$ and $\tanh(4) = 0.9993$. We duplicated the input, and the output barely changed. Thus, also the gradients in this activation range are minimal. In our windturbine control scenario, maximum actions however are beneficial in many scenarios, so we expect actions of 1 and -1 quite often. This trains our net towards the flat outside areas of the tanh activation, and increases our vanishing gradient problem. We solve this by multiplying a constant factor to our activations, moving 1 and -1 further to the inside of the activation. If for example we chose 1.5 as a factor, our last layer activation was $a = 1.5tanh(x)$ where x is the weighted and biased output from the last layer. This allows our actor to choose actions slightly outside of the action range, which has a disadvantage: Q will not see any actions outside of $[-1, 1]$ during Q training, as the training samples to Q are clipped. Thus, Q has to provide estimates for values outside of the area it has seen. As the authors in [FMP] have described, Q tends to overestimate values it has not seen. In fact, when implementing it as described, we did observe the policy to very likely predict actions outside of the action space. We did not manage to fix this problem in the scope of this work, but we could imagine a follow up work where we change the policy gradient equation to account for this overestimation. Alternatively, we could imagine as well to train the Q network on actions slightly outside of the action space with negative estimates to prevent this effect from happening.

In the end, we could not find a definite solution to our vanishing gradient problem and thus had to stick to the smaller architecture, loosing the ability to learn more complex tasks.

4.6 High action gradients

We observed that our controller delivers high action gradients and reacts strongly between time-steps, sometimes jumping from no pitch/torque all the way to the maximum. This happened especially in the beginning of a training run, as the relations between input and output were still very random and sudden spikes not yet trained away. This creates vibrations in the structure and especially high blade pitch changes lead to blades breaking or falling off. Even worse, QBlade was not able to handle this type of structural failure and completely crashed in these scenarios. We tried different methods to circumvent this.

4.6.1 Gradient actionspace

At first, we implemented gradient actionspaces, where the controller output is not an absolute pitch or torque value but a difference to the last output, starting with everything set to 0. For aiding training, we added the actual actions taken to the observation space of the next step. This way, we could clip the action gradients by reducing the gradient action space, so that in each time-step, only a fraction of the actual action space could be traversed. If the controller would output a positive change at maximum pitch/torque or a negative change at minimum pitch, this would be ignored. This did effectively limit action gradients, but also did not lead to any sensible results, so we deactivated this again.

4.6.2 Feeding past time-steps

Another idea was to feed the last n time-steps concatenated to the current observations. We hoped that this way, the agent could derive its own gradients internally. Also, a PID controller has a view of the

past. With this extra information, the controller could be able to generalize further and perform better especially counteracting resonant vibrations. The result of this however were high and very noisy Q-losses and Q loss explosions, so we disabled this again.

4.6.3 Clipping action gradients

We eventually solved the problem with the high action gradients by still letting the controller output absolute actions across the entire action spaces, but if gradients exceed a certain threshold, the actual action taken is set to the nearest value with sensible gradients. We defined the sensible gradient vector g according to [JBMS, Table 7-2]. The clipping function on action a and previous action a' can be defined as $\text{clip}_a(a, a') = \min(\max(a, a' - g), a' + g)$. The results of this are again clipped to be in the action space. Though the actions returned from the policy are still wildly unstable, they at least didn't result in death of the turbine anymore and still we didn't hinder convergence through gradient actionspace such as in section 4.6.1. With this, we could completely exclude QBlade crashing due to unrealistic pitch changes.

We at first hid this clipping from the replay buffer, but then decided to store the actual values taken, as the models then don't have to also learn the clipping effect. Before we reasoned, hiding this clipping effect would result in a more efficient exploration of the action space, as the policy can predict all actions without restriction and thus easily sweep over the action space. However, we observed better learning when not hiding the clipping from the replay buffer.

4.6.4 Pretraining the policy

We added a small period of direct pretraining of the policy on the actions taken by the expert policy during random exploration. We hoped for a more sensible default policy at the beginning. We summarize this in algorithm 3, parametrized with training time $T = 10000$ and batch size $N = 32$

Algorithm 3: Policy pretraining

Use prefilled replay buffer $\mathcal{R}$ from expert exploration
Use expert policy $\mathcal{C}$ from expert exploration
for *t = 1, T* **do**
  Sample a random minibatch of N states (s_i)
  Calculate noise-free expert actions $a_i = \mathcal{C}(s_i)$
  Update the actor by minimizing the loss: $L = \frac{1}{N} \sum_i (\mu(s_i) - a_i)^2$
end

The training is relatively straightforward. We could see a better performance of the policy in the first steps in contrast to not using pretraining, but the effects were minimal. So when we decided to abolish expert exploration, we also abolished this.

4.7 Other improvements

Because still generally not performing well, we added some techniques which we thought could aid general performance.

4.7.1 Prioritized experience replay

[SQAS] proposed a method to sample experiences from the replay buffer according to how much they benefit training, and not just uniformly. This method is called Prioritized Experience Replay (PER). According to how much they benefit training means that we add a priority to the experiences, which we somehow derive from the performance of the algorithm on that samples. In their paper, they showed how especially cliffwalk problems, but effectively all problems, benefit from using prioritized experience replay. We argue that our windturbine is similar to such a cliffwalk, as random exploration usually only shortly passes sensible operation range and then quickly destroys the turbine. Seeing this short high reward might not be enough for the controller to learn that being there is good. Furthermore, the technique is generally promising as in Rainbow-DQN [HMvH$^+$], PER made up for most of the gain on the algorithm. In the original case, they use it on DQN, but we will apply it to DDPG. There hasn't been much work on how to use PER in conjunction with DDPG, in fact we only found two papers: [HZ] and [ZLZH]. In the rest of this section, we will at first present the general prioritized experience replay algorithm, then we will discuss the papers and then present how we incorporated PER ourselves.

As knowing how much a sample benefits training is difficult, [SQAS] propose to approximate this importance by the Temporal Difference Error (TD-Error) $\delta = y - Q(s, a)$. This difference, which in Q learning is comparable to the loss of the approximator, tells us how far off the prediction was in this step. So, alongside with the experiences, a sampling priority for each item $p_i = |\delta|$ is stored in the replay buffer. In theory, we would have to calculate this TD-Error every time we update our critic. This would mean to feed our entire replay buffer through the critic on every update step. Clearly, this is a huge performance hit, which is why in the algorithm, they approximated the real errors by only updating the error on samples that are anyway being used for training. We calculate the TD-Error anyway for the critic update, so we can store this without performance penalty. There is a risk though, that older experiences which used to produce low errors but under a new critic produce high ones are not being fed soon with this approximation. New samples are always stored with maximum priority to make sure they are sampled at least once.

When replaying, the sampling probability of a sample is set to $P(j) = \frac{p_j^\alpha}{\sum_i p_i^\alpha}$, α being a hyperparameter allowing to fade between uniform sampling ($\alpha = 0$) and pure priority based sampling ($\alpha = 1$). Additionally, usually a small ϵ is added to avoid zero priorities: $p_i = |\delta| + \epsilon$. Because we introduce oversampling bias by sampling some experiences more often than others, a correction term is applied to the TD-Error: $w_j = \frac{(N * P(j))^{-\beta}}{\max_i w_i}$. It is simply multiplied to the real TD-Error. β is a parameter describing how strongly to correct oversampling, where $\beta = 1$ is full correction and $\beta = 0$ no correction at all. This importance sampling is a common technique in reinforcement learning and we have used it before in our background part in equation 3.10.

In addition to sampling directly based on $|\delta|$, [SQAS] propose a variant where they sample according to the rank of an experience: $p_i = \frac{1}{\text{rank}(i)}$ where rank is defined as the position in the replay memory when sorted according to $|\delta|$. This method is more robust to outliers and slightly increases performance, but because of the extra computational effort of having to calculate the rank every time, we decided to implement the first method. Also, [ZLZH] measured worse performance of the rank based version in contrast to even plain DDPG.

The importance sampling is also the reason why we deem it not so trivial to introduce PER to DDPG, as there is no specified way how to integrate this into policy updates. But first we will describe two other efforts. [HZ] implemented it straightforward on the Q-Learning part of DDPG and ignored any effects on policy updates. They measured improvements tackling the OpenAI Pendulum task in contrast to vanilla DDPG. However, they only evaluated this pendulum task, which we already suspect not to be representative of our windturbine problem. We prefer to rely on the work of [ZLZH] - the team has formulated experience replay as a learning problem and they trained another prediction network on which replay to choose. They evaluated their learnt experience replay against sampling based on rank or

priority and found their algorithm to outperform both, while rank-based sampling performed far worse than even vanilla DDPG and priority based sampling slightly better than vanilla DDPG. However, the team completely disregarded importance sampling in their work. Also, we deem it as unnecessary to introduce yet another learning task, especially because we are calculating on weak hardware.

Because neither of the works seemed solid enough for us to implement it, but because the learning improvements in Q-learning are very promising, we decided to create our own mixture of PER and DDPG. We decided to implement the Q-Learning part straightforward as in [SQAS] and including importance sampling. We used the reformulated TD-Error of DDPG with the target actor and critic. To avoid oversampling bias in the policy, we only applied PER-sampling to the training part of the critic and trained the actor on uniformly sampled data. It can be argued, that as the policy μ is anyway being trained on the state density function of another policy ρ^β, it can deal with that oversampling bias. However we doubt adding more bias is beneficial to the training, so we decided to sample uniformly for the policy.

Because we saw performance problems storing the priorities alongside the experiences and directly sampling from a 100000-item weighted array, we followed the advice of PER to use a sumtree datastructure, a binary weighted tree in which each node holds the sum of weights of the children. At first we implemented sampling from this as a tree walk, however found that this way we are in danger of sampling the same item several times. In edge-cases, we ended up with a batch consisting of only one item repeatedly in a Q loss explosion, which resulted in batch normalization variation compensation to perform a divide by zero, storing NaN in our weights and effectively ruining the network. We replaced it with a draw without duplicates. Also, we did not store probabilities right away but kept the unnormalized experiences $p_i = (|\sigma| + \epsilon)^\alpha$ and only normalized them to probabilities right before sampling.

4.7.2 Data augmentation

We still had some parts of the code left where there wasn't any noise involved, so we decided to add some to the training process. Data-augmentation with noise is a common technique when using neural networks [Bis, p.347] to reduce overfitting, so we expected it to also make our model generalize better. There have been some experiments with data augmentation in reinforcement learning [CKH$^+$], but it is not often mentioned on tutorials or blog posts. We implemented adding a small noise term to the sampled states, actions and next states (s, a, s') when performing experience replay. We didn't really see any indication of better or worse generalization, but as implementing it wasn't a big issue and we neither saw nor could imagine any disadvantages, we still added this feature and set the noise level to a very small value.

[FvHM] addressed something similar on actions generated by the target policy, see section 4.8.3 for this. They argued that noise added to the target action smoothes out the critic. With the same argumentation, also our replay noise would smooth out Q estimates.

4.8 Exploding Q-Loss

We observed that our Q-Loss, especially after a death, explodes into e15 magnitude values, while most of the times it stayed around e2 magnitude. Instable Q-Losses are a general problem with DQN and DDPG, however literature normally doesn't speak about Q loss explosions, or if yes than in the form of divergence. It is possible that a high learning rate makes the Q function jump away from the target, stepping over the minimum of predicting target Q values and instead predicting higher, then lower, then even higher values. This process however is different to what we observed, as during our Q loss explosions mean Q estimates did not jump as well. Also, the loss usually jumped up within 1 to 10 training steps and then decreased again, in contrast to a divergence which would somehow constantly and smoothly escalate. So we tried to figure out what caused our loss explosions.

4.8.1 Huber loss

The DQN authors are using Huber-loss [Hub] for the critic instead of mean squared error loss as described in DDPG paper. In fact, the authors of DQN have caused some confusion on this in the RL-community. In their journal paper [MKS+b], they write

> We also found it helpful to clip the error term from the update [...] to be between -1 and 1

which could be interpreted as loss clipping. However later they explained that they actually switched to mean absolute error instead of mean square error outside those boundaries. This is the exact definition for Huber loss.

$$\text{loss} = \frac{1}{N} \sum_{i \in N} \begin{cases} \delta_i^2 & \text{if } \delta_i <= 1 \\ |\delta_i| & \text{else} \end{cases} \tag{4.1}$$

Although loss explosions got less frequent and smaller in magnitude with this update, we still saw some. In general we could observe an improvement in training, so we kept this active.

4.8.2 Large batches

We also had to increase batch size up to 128 to counteract noisy observations. A large batch size results in less noisy losses, as the loss is averaged across a high number of samples. However, large batch sizes can also limit learning performance, as the network is smoothing across a high number of samples. Generally, batch sizes of 32 or 64 have established in deep learning, but we could reduce our Q loss problem slightly by taking a bigger batch size.

4.8.3 Double critics

We also had a look at [FvHM]. They propose a set of three mechanisms that address function approximation error in DDPG, and they call it Twin Delayed Deep Deterministic Policy Gradients (TD3). At first, they show in their section 4.1 that Q learning with a deterministic policy tends to overestimate true Q values. This has been observed in normal Q learning as well, but hasn't been generalized to DDPG before. They show, that the interaction of the policy and the Q network leads to a residual error accumulating over the recursive nature of Q updates. To mitigate this, they at first proposed to use two Q networks and choose the minimum of the two to compute the target y. They combine this with the target network idea, so effectively we end up with 6 networks now. They called this technique *Clipped Double Q learning*. We illustrate this in figure 4.1. The Q target is computed as

$$y = r(s, a) + \gamma \min_{i \in 1,2} Q'_i(s', \mu'(s'; \theta^{\mu'}); \theta^{Q'_i}) \tag{4.2}$$

As they observed that a quickly changing policy introduces instability into the Q learning process, they proposed to update the policy less often than the actor. They do this simply by updating the policy only when $t \mod d = 0$ where t is the time-step and d describes how many time-steps to wait before an actor update. They call this *delaying policy updates*. Lastly they proposed adding noise in the target policy calculations to achieve smoothing around action value estimates. So they changed the target updates again to:

$$\tilde{a} = \mu'(s'; \theta^{\mu'}) + \text{clip}(\mathcal{N}(0, \tilde{\sigma}), -\tilde{c}, \tilde{c}) \tag{4.3}$$

$$y = r(s, a) + \gamma \min_{i \in 1,2} Q'_i(s', \tilde{a}); \theta^{Q'_i}) \tag{4.4}$$

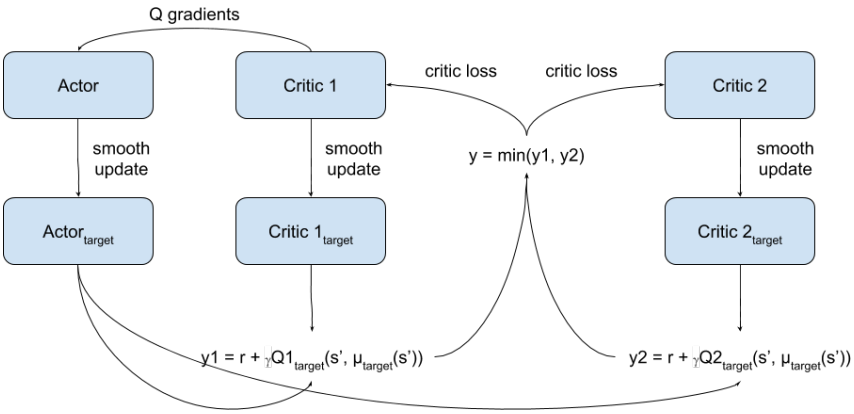


Figure 4.1: Clipped double Q learning

Where $\tilde{a}$ is the target action with added noise on it. In their appendix they describe that they furthermore clip this target action to sensible action range. This is called *target policy smoothing*. Summarizing, TD3 consists of three mechanisms, Clipped double Q learning, delayed policy updates and target policy smoothing.

When implementing this, we found that we now have two TD-Errors which left us with a choice how to implement PER in conjunction with this. We could either use one of the two errors. When implementing this, we found that the critic whose error we chose for PER performed way better than the other, effectively rendering it useless to have two critics. We also tried to use the maximum and minimum of the two, which yielded more stable results, where minimum seemed better than maximum. Finally, we ended up with adding the two errors together for prioritization.

However, we could not get TD3 to work properly, with or without PER, and didn't see any convergence. We thus disabled all three features, but found that using a lower learning rate for the actor has a similar stabilizing effect as actor delays. Also, we later found that target policy smoothing helps with a problem in death conditions, but we didn't have it active for most of the training.

4.8.4 Regarding death conditions

We were observing many burning turbines and flying-off blades in our experiments, so we had a look at how DDPG handles death conditions. In DQN, death conditions are accounted for by changing the Q target in the terminal step to be $y = r$ instead of $y = r + \gamma \max'_a Q(s', a')$. In fact, we regard the lack of this in DDPG as a bug. To explain why, let's have a look at our Q-function again:

$$Q(s, a) = r(s, a) + \gamma Q(s', a') \tag{4.5}$$

The right term of this equation $\gamma Q(s', a')$ is likely several magnitudes higher than the left term $r(s, a)$. With our γ = 0.99, circa 99 times as high as $\int_1^\infty 0.99^x \approx 98.5$. Though we punish death conditions severely, a death condition will not exceed 99 times the rewards of normal operations without causing a Q loss explosion. So if, after a death condition, a good condition is reached due to resetting the environment, our algorithm would gladly take that penalty. For this reason we introduced the updated death notion

$$Q(s, a) = r(s, a) + \begin{cases} 0 & \text{if death} \\ Q(s', a') & \text{else} \end{cases} \tag{4.6}$$

This update notion is similar to the update notion from DQN. So now, instead of $y = r + \gamma Q'(s', \mu'(s'))$ we change it to be $y = r + \gamma Q'(s', \mu'(s')) * (1 - d)$ where $d \in 0, 1$ symbolize the death conditions.

We believe this addition to DDPG to be novel. Also, in section 6.2, we discuss why it is necessary to also make sure the algorithm sees mainly positive rewards. Unfortunately, we didn't know this at the beginning of our experimentation phase so this is not included in our original algorithm.

4.8.5 Clipping observations

Finally, we found that the solution to our problem lay in QBlade. In certain conditions, QBlade returned observations 20 orders of magnitude away from normal operation for a single timestep. As we based our reward calculation on the observations, the reward was also far off from normal reward. This condition happened only every 200k steps or so, and as we did not log every time-step, these spikes didn't end up in our logs. We decided to filter out these spikes by clipping rewards and observations after normalization to $[-3, 3]$. We reference this clipping by clip_s.

4.9 Concluding all changes

We want to quickly summarize all the changes we made. We tried several things and abolished some of them because they didn't show results. Among the ones abolished, we had *parameter-space noise* due to it not giving a visible advantage. We abolished *expert exploration* and *policy pretraining*, because adding human knowledge to the training process was what we wanted to avoid in the first place. The advantage of having them was small anyway. We tried predicting *action gradients* instead of actions directly, and stopped because results were poor. *Feeding past timesteps* caused even more instable Q losses so we didn't use this either. Also, we couldn't get *TD3* to run as a whole. *Last-layer activations* drifted our policy out of the action space so we deactivated this as well.

We were unsure about the effects of some mechanisms. *Data augmentation* has neither proven helpful nor limiting, but because we think it is a good thing, we left it in with small magnitude. Same goes for *decaying action noise*. We could get *target policy smoothing* to work but only used it on our last experiments.

Other mechanisms have proven helpful. *Clipping action gradients* was necessary to stop the agent from destroying blades by too quick pitching movements. *Clipping observations* helped removing spikes from faulty simulation results. *Large batches* and *Huber loss* added stability to the Q function. A *simpler architecture* prevented zero gradients.

Lastly, we have three mechanisms which are of higher complexity. *PER* is uncommon to DDPG, but it has delivered better results for us. *Normalization* was critical for learning anything at all and removed a lot of our problems. And lastly, we argued why *death conditions* are not properly handled in DDPG and proposed a solution.

5 Algorithm

After our experimentation phase, we will present our algorithm, on which we perform our evaluation. At first, we describe how we encapsulated the QBlade environment. At second, we will present our replay buffer, our network structure, our exploration strategy and our update algorithm.

5.1 QBlade

As mentioned before, we set the yaw angle to always 0 degrees, as in reality controlling the orientation of the nacelle is trivial and in our simulations, wind will always come from the front. The generator torque on our NREL 5MW turbine has a sensible range of 0 to ca 4.6×10^6 Nm, we copied exact value from [JBMS, Table 7-2]. For this it should be noted that the torque in the specification sheet is measured after a gearbox, while qblade needs torque before the gearbox. Thus the gearbox ratio of 97 has to be multiplied to the values in the specification. Because changing torque from zero to maximum in one time-step is not realistic for a real turbine, we restrict it to move a maximum of ca 1.5×10^6 Nm per one time-step of 0.1 seconds. We again took exact values from [JBMS, Table 7-2]. If our agent chooses a value outside of that area, we set it to the closest value inside of that area. Blades can be pitched between 0 and 90 degrees, 0 being not pitched out at all and operating at maximum efficiency and 90 resulting in no aerodynamic torque from the rotor whatsoever. A sensible pitch motor can only turn the blades at a limited speed, and the NREL specs give a maximum change of 8 degrees per second. We again clipped controller inputs to a sensible value inside that. The simulation theoretically accepts a full pitch change in one time-step, however as inertia on the blade is so high for such a change, the blades break instantly and the rest of the simulation needs to be reset.

We observed problems with our controller when reaching extreme limits of the simulation in the course of our experiments, thus we decided to reset the simulation at a rotational speed of 3 rad/s, as our turbine normally operates at ca 0.8 rad/s. Also, high generator torque inputs cause the simulation to rotate the rotor backwards with a negative energy yield, effectively creating a multimillion-dollar leafblower. As this is neither a realistic scenario, we also reset the simulation at negative rotational speeds. Also, power outputs over 10MW and excessive swinging more than 2 magnitudes higher than normal vibrations are considered a death condition.

When resetting the simulation, we artificially stabilized it to 0.8rad/s rotational speed with the precomp option in QBlade. During this stabilizing phase, torque and pitch are fixed to 0. After vibrations at simulation start were overcome, we handed over control to our agent and ended the precomp phase. Starting with an uninitialized simulation with standing rotor very likely enters small negative rotational speeds at the beginning and thus results in constant resetting.

We hid all observations except for power and rotational speed from the algorithm and also merged all pitch controls into a single variable, implementing collective pitch control.

5.2 Core algorithm

We implemented a replay buffer with capacity 1e6. To enable PER, we needed to sample according to priorities proportional to a metric, in our case absolute TD-Error. We used PER with minimum priority $\epsilon = 1 \times 10^{-6}$, prioritization intensity $\alpha = 0.5$ and importance sampling correction $\beta = 0.5$.

We implemented our actor model according to the DDPG paper, but with less neurons per layer. Instead of 400 and 300 neurons, we use 32 and 16 neurons. Our critic is also according to the DDPG paper but with less neurons. Instead of 400 and 300 we use 64 and 32 neurons. We use batch normalization between the layers. All states and rewards seen by the network are normalized to be between $[-1, 1]$ 90% of the time and are clipped to $[-3, 3]$. Actions are normalized to $[-1, 1]$ and clipped to not extend sensible action gradients. Learning rate for both actor and critic are 1e-4 and we use adam optimization. We train with batch size 128, discounting factor γ = 0.01 and target update rate τ = 0.05. Weights are initialized by uniform random in $[-0.5, 0.5]$.

Critic updates are done similar to DQN updates with PER. We use importance sampling to correct for PER overestimation bias in the critic and redraw samples uniformly for policy training. We regard death conditions in the Q update as in DQN. Also, we use Huber loss instead of MSE. All in all, algorithm 4 summarizes all improvements:

Algorithm 4: Our DDPG algorithm

Randomly initialize critic network $Q(s, a; \theta^Q)$ and actor $\mu(s; \theta^\mu)$ with weights θ^Q and θ^μ
Initialize target networks Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer $\mathcal{R}$
Calculate normalization factors based on data from a previous run
for *epoch = 1, M* **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for *t = 1, T* **do**
    Select action $a_t = \mathrm{norm}(\mathrm{clip}_a(\mu(s_t; \theta^\mu)) + \mathcal{N}_t)$ according to the current policy and exploration noise
    Execute action a_t in ε and observe r_t, d_t and s_{t+1}. Clip and normalize rewards and states according to clip_s and norm
    Store transition with highest priority $(s_t, a_t, r_t, s_{t+1}, d_t, \max_i(p_i))$ in $\mathcal{R}$
    Calculate sampling probabilities $P(i) = \frac{p_i}{\sum_j p_j}$
    Sample a random minibatch of N transitions $(s_i, a_i, r_i, s_{i+1}, d_i)$ from $\mathcal{R}$ according to P
    Calculate importance sampling weights $w_i = (N \times P(i))^{-\beta}$ and normalize $w_i = \frac{w_i}{\max_j w_j}$
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}; \theta^{\mu'}); \theta^{Q'})(1 - d_i)$
    Calculate TD-Error $\delta_i = y_i - Q(s_i, a_i; \theta^Q)$
    Store TD-Errors in the replay buffer: $p_i = (|\delta_i| + \epsilon)^\alpha$
    Apply importance sampling $\delta_i = \delta_i w_i$
    Update critic by minimizing the Huber loss: $L = \frac{1}{N} \sum_i \begin{cases} (\delta_i)^2 & \text{if } \delta < 1 \\ |\delta_i| & \text{else} \end{cases}$
    Sample a random minibatch of N states (s_i) from $\mathcal{R}$ uniformly
    Update the actor policy using the sampled policy gradient:
    $\Delta_{\theta^\mu} J \approx \frac{1}{N} \sum_i \Delta_a Q(s_i, a; \theta^Q | a = \mu(s_i)) \Delta_{\theta^\mu} \mu(s_i; \theta^\mu)$
    Update the target networks:
    $\theta^{Q'} \leftarrow \tau\theta^Q + (1 - \tau)\theta^{Q'}$
    $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1 - \tau)\theta^{\mu'}$
    On death condition, reset the environment
  end
end

6 Evaluation

For the evaluation, we used a machine with an Intel® i7 4-core processor, 16GB RAM and a Nvidia 1080 GTX with 8GB as accelerator. On this system, a fully realistic run with 300k steps would take roughly 200 hours, of which almost everything would be used by QBlade. Partly this is because we didn't manage to enable hardware acceleration because of compilation problems, but computing a fluid simulation is always time-consuming. We could bring the computation time down to 12 hours per run by deactivating wake calculations and using a different structural computation method in QBlade. Wake calculations are responsible for calculating the effects of the rotor on the surrounding air. Without this effect, the rotor can draw infinite energy from the air as it can generate lift without slowing down the air. Effectively, this increased instability of the system, as now the rotor does not get slowed down at high rotational speeds, but otherwise we would not have been able to evaluate our work. Also, during most of our training we were able to fit a single run into the 16 GB RAM of the machine, but after optimizing our code for memory usage, we could accommodate 3 easily or 4 with heavy swapping penalties on one of the runs.

We ran every experiment 4 times except stated otherwise. We visualize our measures as confidence plots, for which we calculated the mean and a 90% confidence interval over the results. The mean is displayed as a solid line, while the confidence is plotted as a colored background.

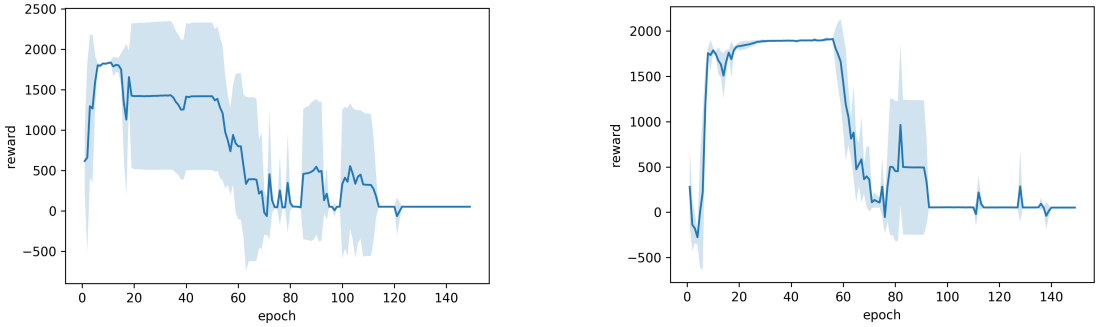
Unfortunately we couldn't evaluate all our improvements in a proper grid search and measure their impact. We will limit us to very few scenarios, but hope to discuss them thorough enough to provide valuable insights.

6.1 Hold speed

We start off with the simplest task. The task to solve was to hold a rotational speed of 0.8 rad/s over the time of 2000 steps, starting with a simulation that has artificially been stabilized at the desired rotational speed with 0 degrees pitch and 0 generator torque. Wind-speed is kept constant at 11m/s. The agent gets a reward of 1 per step if it exactly matches rotational speed, deviation from it is linearly punished, 0 rad/s would yield a reward of 0. Thus, the theoretically possible maximum reward would be 2000, holding the turbine at perfect speed for the entire length of the simulation. However, as we start with 0 pitch and torque, action gradient limitations will naturally incur a small penalty at the beginning as these values need to be adjusted to sensible values and the resulting rotational speed deviation needs to be recovered. A perfect policy would thus reach a score of a litte bit below 2000. This task is simple because it can be achieved either through pitching out, increasing generator torque or a combination of both. This task is not realistic, as it can be solved without generating energy whatsoever.

We look at two different learning rates for the critic network, 1e-3 and 1e-4. The latter is what was recommended in the DDPG paper.

Our results are plotted in figure 6.1. Each of the controllers reaches close to theoretical optimum and is able to hold it for a number of epochs. With learning rate 1e-3, we reach it somewhere around epoch 10, with learning rate 1e-4 a bit later around epoch 25. We also tried 1e-5, but we could not see any convergence with this, so we omitted the plot. The time we are holding it also raises with learning rate. The most aggressive learning rate reached 1828.8 average epoch reward between epoch 7 and 11 and a single maximum of 1917.5 epoch reward. On the lower learning rate, it held an average of 1896 points



(a) critic lr=1e-3 (b) critic lr=1e-4

Figure 6.1: Learning rate comparison for the critic network

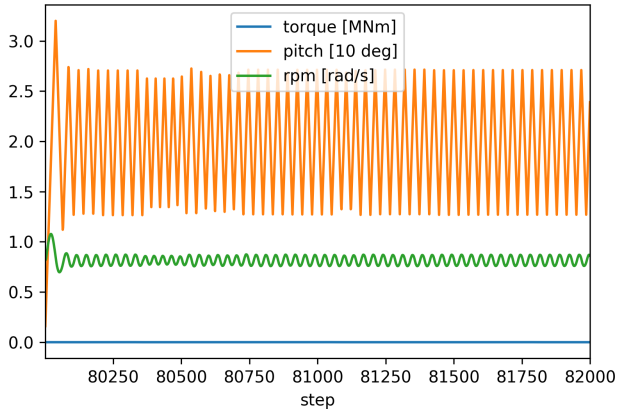


Figure 6.2: Pitch, torque and rotational speed from a sample epoch

from epoch 30 to 56 and reached a maximum of 1968.4 points. The maximum we could reach with our hand-designed PID controller was 1981 points, where the 19 points were lost through the swing at the beginning and then rotational speed was held perfectly. A totally random policy reached 440 points on average.

So, we were able to reach sensible control for a number of timesteps. 4 epochs on the high learning rate and 26 epochs on the low one ranged in an area which is close to optimal. We widely outperformed the random policy and also we needed quite some time of trial and error until we were able to beat our learnt policy in this task with our handmade controller.

Interesting to see is the strategy which the agent found. As described before, both increasing pitch and increasing torque reduces rotor speed, so through both it would be possible to control the rotational speed. However, all of our runs decided to implement pitch-only control and kept torque constant at zero, as possible to see in figure 6.2. A possible explanation for this decision could be that with high pitches, lower forces act on the blades, thus rotational speed is easier to control. However, this insight is above our human knowledge of windturbines. Furthermore should be noted that rotational speed is still oscillating between 0.75 and 0.85 and the pitch is ranging from 15 to 25 degrees (note the scaling of the pitch and torque). In a realistic scenario, oscillations of this magnitude would likely destroy the turbine very soon, as pitching the blades around this quickly creates strong material stress. This makes the algorithm not usable as it is.

Both algorithms converge to doing nothing after a while, which indicates a divergence in the algorithm. Interestingly, we can not see our divergence in the loss. A divergence caused by a too high learning rate would cause a phenomenon of jumping across a minimum, taking gradient steps greater than the width of the local minimum and skyrocketing both the loss and Q estimates, and we observed plenty of these scenarios during our experimentation phase. Here, however, the Q loss stays constant on such divergence,

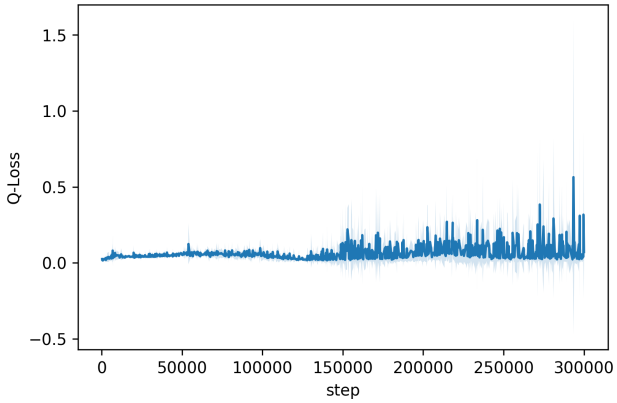


Figure 6.3: Critic Q loss for lr=1e-3

as plotted in figure 6.3 for the diverging run (a) with the high learning rate. So, what we have is not a traditional Q divergence. Instead, we must have learned some maximum in the Q function, towards which the policy now optimizes. This minimum is doing nothing and leaving the turbine still. In this minimum, death conditions are strongly reduced in number, as turbines tend to not die when they don't rotate. This minimum could be learned after learning the maximum towards which the policy converged before but on samples generated before the policy was able to reach this maximum. Learning about the negative nature of death conditions is a slower process, as death conditions in our dataset are relatively sparse, so learning the penalty happens after optimizing the policy towards a good behavior. With the lower learning rate 1e-4, this happens even later. With even lower learning rates however, we have accumulated enough death condition for Q to already learn them before optimizing the policy towards a sensible state, so a learning rate of 1e-5 does not bring us to any convergence (not plotted here). A solution for this would be to train the network for a longer time until the death conditions from the beginning ended up outside of the replay buffer. When that happened, movement towards higher rotational speeds could be rated as actually viable by Q without Q already punishing the near certain deaths of the policy. Training for longer however alleviates our problem of vanishing gradients, as the policy is pushed towards actions at the borders of the action space when bringing the turbine to a standstill. Training for a longer time, we observed very small gradients. We don't know whether it would recover from the low gradients, and trying is outside of our computational possibilities. We will call the process of converging to doing nothing and strictly avoiding any deaths *panic mode* in the rest of the evaluation.

From this we see that higher learning rates cause a more stable learning, but reach convergence later. The results suffer from strong oscillations with both the low and high learning rate. We do not expect the Q function to learn resonant behavior of the windturbine in the few training steps which we allowed it to run, so we are not surprised that the policy can not yet counteract this. In a production systems, oscillations of the magnitude which we observed would wear the turbine strongly and would likely destroy it within a few days.

6.2 Hold rated power

A more difficult task to solve, but also a more realistic one it to hold a rated power. Analog to our hold speed task, the agent gets a simulation which was artificially stabilized at 0.8rpm with pitch and torque 0, and should generate exactly 5MW of power during an epoch of 2000 steps. This power is the reference output of the NREL 5MW turbine, and in our wind scenario it should theoretically be able to reach this power. Wind speed was kept constant at 11m/s. Again, the agent gets a reward of 1 for hitting 5MW perfectly, and 0 for having no power at all, so a perfect policy would get something a little below 2000. Unfortunately, we did not get any policy to achieve something close to optimal, the best

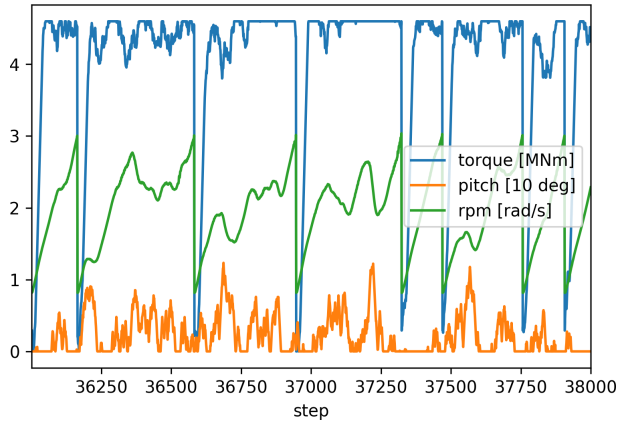


Figure 6.4: An epoch from the hold power task

score we achieved was 1500 points with an average of 5 kills per epoch. Unfortunately, this result wasn't very stable, so we can not plot an average over many runs. An example for an epoch with high reward is figure 6.4. Here, the controller maxed out torque and tried to hold the turbine stable with pitch. It reaches rotational speeds above 2rad/s though. Power output for these insane regions was up to 9MW, which would have instantly roasted any electronics in a real wind turbine. Also, blades would have taken serious damage on these rotational speeds.

At the end, most controllers converged to leaving pitch at zero and left the turbine to die regularly with maximum generator torque. So again, the controller has learned something which obviously isn't an optimum in the global reward function. We again don't see divergence indicators, so we need to ask again how this relatively poor strategy became an optimum in the Q estimate.

A possible explanation for favoring deaths could be that our death penalties are spikes in the reward functions. As explained in section 4.3, we add a constant penalty upon death but not in the steps before. As our Q function approximator approximates smooth functions, it will likely have a high TD-Error on these samples. This in turn will cause PER to replay these samples relatively often in comparison to other samples. As we didn't use a PER beta of 1 to fully correct against importance sampling bias but 0.5 as proposed in the paper, we could imagine that still some oversampling happened along these points and they actually ended up being largely overestimated.

Also, we could imagine that the Q function gets deformed before deaths and maybe actually exposes a local maximum right before it. The spikey nature could lead to weird situations in a high-dimensional regression, which could be similar to overfitting. As the regression wants to fit to the outlier, it might create a strong curvature before this. To illustrate this, let's plot a typical overfit scenario in figure 6.5. The dot on the bottom right is our spike in death. Deriving along the linear regression (blue line) would lead us away from the death. The orange line however exposes a local maximum shortly before the death. If our policy for some reason got stuck in this local maximum, it would be trained to stay close to a death condition, which would, because of system instability, lead to many death conditions happening. Oversampling effects might then even increase this phenomenon.

However, there is another, more credible explanation to the problem. If you remember section 4.8.4, we explained how in the Q update, the advantage of a single result is roughly 100 times lower than accumulated future rewards in the update notion to Q: $Q(s,a) = r + \gamma Q(s',a')$. There is a case in which it could be advantageous to die. Let's suppose we have a terrible policy which very likely produces rewards below zero, or an environment in which it is difficult to obtain rewards greater than zero. Thus, our Q term would likely be negative, as however we act, we will incur negative rewards. As discussed before, Q is likely of high magnitude, so eliminating the high negative Q for a 0 would be a good strategy. A policy might learn this strategy and result in a behavior similar to what we observed. Thus we need to make sure the Q term is never negative. We propose a simple method, we circumvent this by adding

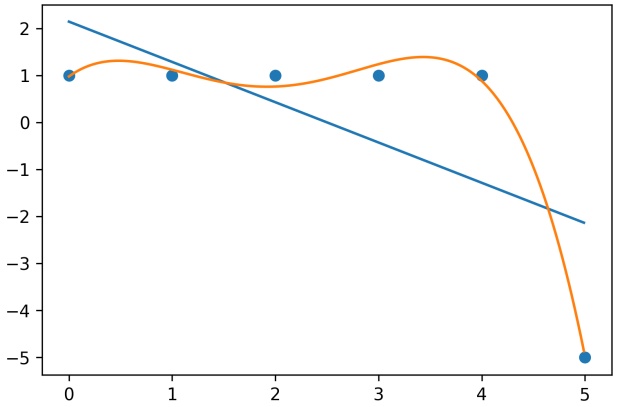
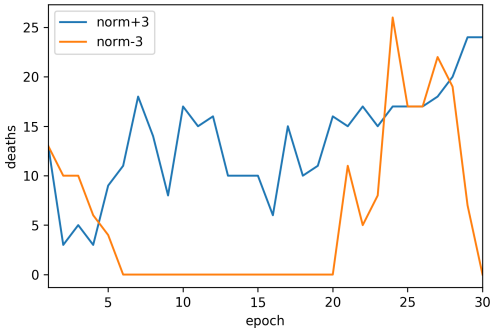
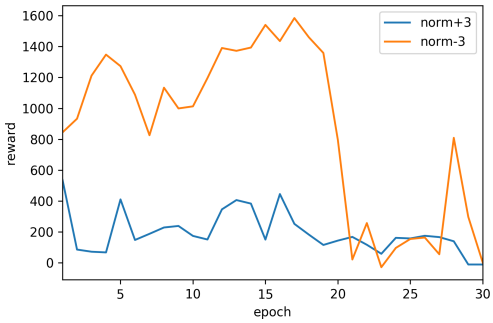


Figure 6.5: Overfitting a 5 point regression



(a) Number of deaths per epoch



(b) Cumulative reward per epoch

Figure 6.6: Comparing adding and subtracting a constant factor to rewards

the lower clipping bound to rewards after clipping them. This results in the policy seeing only positive rewards, which would cause Q estimates to also be positive.

Our environment is in fact an environment in which is is difficult to achieve positive results. In contrast to the hold speed task, a small change in rotational speed does not end up in a small change in reward, but the turbine is required to first spin the turbine up to cut-in speed and then torque the generator up. This results in the average Q estimate to be below zero, as executing this chain of actions happens far less likely than to for example kill a turbine. When plotting average Q estimates (not shown here), we in fact see negative Q estimates.

We evaluate whether our explanation with the negative Q parts holds by running once with rewards in the range of $[0, 6]$ and once rewards between $[-6, 0]$. We see the result of the direct comparison in figure 6.6. Unfortunately, we can only perform a single run of 20 epochs for evaluating these theories due to time constraints. To improve performance, we performed 2 training operations per step, so epoch 20 is comparable to epoch 40 from our previous setting. In our results, we can see that the controller which gets only positive rewards performs widely better, and nearly without killed turbines. Both controllers fail and converge to doing nothing then. Because of our low number of runs and short evaluation time, we only see this as a hint to support our thesis, not a full proof.

We do not evaluate our spike theory again but instead run a training on the hold power task with our new normalization addition.

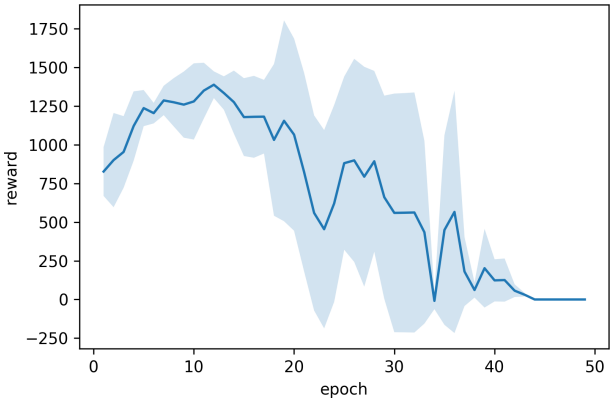


Figure 6.7: Results of the hold power task

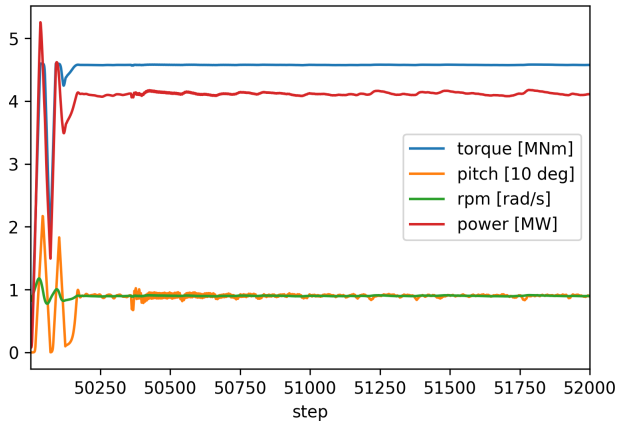


Figure 6.8: An epoch from the hold power task

6.3 Hold rated power with death conditions

We did a last evaluation run similar to section 6.2, however with the aim of holding 3.5MW instead of 5MW, as we deemed this task slightly easier. Also we changed the critic size to 128 and 64 neurons because in previous runs we observed going to panic mode quickly. Using a bigger and thus slower to train network made us hope to achieve sensible results later but with more stability.

In figure 6.7 we can see at least some convergence happening. Though no agent reached anything near perfect, we are at least seeing some improvements over a totally random policy, which gets an average reward of 410 points. If we look in detail at actions the controller took, as displayed in figure 6.8, we see that it even managed to smooth out vibrations. However, it stabilized at 4 MW instead of 3.5MW. We would again explain this miss to the early stage in training we are looking at. Also only one of our 4 runs managed to stabilize this well, the others still had severe oscillations occur. This instability in training could be due to the bigger network size, however our better results could also be due to the bigger network size. In general machine learning, training bigger networks is usually more difficult but provides the ability to tackle more complex problems. In supervised learning, overfit is a clear indication of a too large network and can be measured directly by validation performance. As we do not have access to such measures in reinforcement learning, rating generalization is less trivial.

6.4 Pendulum

To exclude the possibility that our implementation is buggy, we decided to run our DDPG algorithm on the Pendulum task again. With our hyperparameters, we could not observe sensible results. Only switching back to the original algorithm of DDPG which we already tried in 4.1, the agent managed to balance out the pendulum. Thus we can not exclude mis-implementation as a bug might only occur with our set of hyperparameters, however we suspect that windturbines are just not very similar to pendulums.

6.5 Discussion

To summarize it, the level of control we have achieved does not outperform any previous benchmark controller. In fact, it would destroy most windturbines relatively quickly and does not yield sensible energy. We have poor training stability, as runs are doing very different things. The best words for our results are that at least our agents learned something. We can outperform random policies easily until panic mode sets in, however this is a very low milestone. We conclude that we are not able to reach sensible windturbine control in the scope of this work.

Commonly, algorithms are evaluated over many different problems. The OpenAI Gym suite is the common suite for evaluating reinforcement learning algorithms, and typical problems in this are balance problems (all cartpole* and pendulum* tasks) and actuation tasks (gripper*, reacher*, cheetah*). Balance tasks could be argued to resemble our windturbine in a way, as we want to balance an output as well. However a windturbine is apparently different enough so we can not optimize it with the same algorithm which optimized the other games. It exhibits resonant frequencies in addition to the optimization problem which the controller needs to keep under control. The multitude of resonating elements on a windturbine results in observations being noisy. And also, wind turbines shatter when exceeding operational ranges, which none of the OpenAI gym examples does. This difference made it difficult to train our networks, and without reformulation of the original algorithm wasn't possible.

We did not fix our panic mode after some epochs neither could we get stable convergence on the more difficult task. We were not able to create a definite solution to our vanishing gradient problem and thus had to train with smaller networks than the original authors. We did not properly evaluate all our changes in a grid search and comparison of the individual effects of our changes, so we can not say which change has which effect. In fact, our entire evaluation shows a total of 14 runs, so none of our claims are statistically significant.

However, we at least achieved some things. We could achieve partially sensible control and could outperform a random policy despite the problems we were facing. When solving especially the vanishing gradient and performance problems, we believe that achieving high quality results is possible.

7 Future work

Just because we didn't manage doesn't mean it's impossible. We can imagine that with some additional work, learning stable policies is possible. We have some ideas how we could achieve this result and what was missing in our work.

7.1 Scale up

It was limiting to perform all our experiment on the tiny machine, not being able to calculate even by far as many iterations as the authors of prominent papers and having very limited ability to try different parameters in parallel. Limiting factor in this is currently the simulation, even with the lowered accuracy, it currently uses 90% of the computation time, while we only use 10% to train our networks. As currently, linux hardware acceleration is broken in QBlade, we recommend any future researcher to fix this and to run future experiments on a more powerful machine. This would also allow efficient hyperparameter sweeps, as we saw that hyperparameter choice greatly affects the outcome of the simulation. Also, it would be interesting to run simulations for a longer time. The authors of TD3 ran their simulations for up to 1 million steps, DDPG 2.5 million steps and DQN 50 million. On easier tasks, TD3 could reach first sensible results after 0.2 million steps. With our computational power we couldn't run anything above 0.3 million steps. But we also chose a less complex architecture which should be faster to train, and did reach first results as early as 20k steps.

With high computational power, we could also evaluate many different wind scenarios and turbine models, in contrast to our evaluation which was done on a single wind speed and turbine type.

7.2 Activation problems

We mentioned some ideas in section 4.5.3 we discussed why changing the policy gradient equation or preventing the Q network from overestimating unseen values in combination with the constant factor on the activations could solve our problem with vanishing gradients. Future work could try one of these methods.

7.3 Validate against OpenAI-Gyms

We did not validate our combination of PER and DDPG through an excessive evaluation. Neither did we validate our addition of death conditions that way. For our task, it gave better results in the first stages of training, but it would be interesting to know whether these results are reproducible and whether our algorithm might even yield better results on general reinforcement learning tasks. Running through the suite of OpenAI Gyms would be an interesting work and could show whether our critique on DDPG was justified. Also, we might be able to get TD3 to work there. This comparison might also yield some insights into what makes our case special in comparison to general reinforcement learning.

7.4 Train PID inputs

As PID controllers have proven well in controlling a turbine and neural nets directly performed poorly, we could imagine to predict PID controller parameters and error targets instead of directly building the controller. Alternatively, feeding the algorithm with derivative and integrative inputs similar to what we described in section 4.6.2 could enable the policy to build more resonant-robust results.

7.5 Reward functions

We did all our training under the simple hold-speed or hold-power reward function. A realistic reward function however would incorporate power generated and long-term damages occurred. Especially the last part is tricky, as it would somehow have to be included into the scalar reward function. An interesting approach could be to try to train under multi-dimensional reward functions. To our knowledge, there haven't been efforts to accommodate this into Q-learning. Also, we could imagine interpolating between different reward functions during training.

To combine the learning ease of a simple reward function with the performance of a complex reward, we could imagine fading over between a simple and complex reward function. In early stages of training, the network is conditioned to only output anything sane, while in later training it could optimize complex goals such as damage prevention more. This could be achieved by increasing the magnitude of the reward function further down in training, like after focusing on the big chunky goals zoom in on more fine-grained problems.

7.6 Expert policy training

We implemented expert policy pretraining. Alternatively, we could imagine leaving the expert policy training active the whole time, not just once after random exploration, and to reduce the learning rate of the expert policy training gradually. This way, in initial training, predicting the expert policies actions would have a higher impact than what the Q-function would suggest, whereas in later training with the smaller learning rate, deviations from the expert policy would be punished less and the policy is allowed to deviate further from the expert policy. This would involve two backward passes per training iteration and would require prediction targets from the expert policy on new seen observations

7.7 Expert policy in instable conditions

When the turbine is reaching a dangerous state (high/low rotational speed, high vibrations), we could fall back to our expert controller to save the situation. This would result in a safe controller, that upon insane policies would fall back to sane behavior. It would degrade exploration to what we deem safe, so a controller with this fallback method would never explore a death condition. This would not necessarily hinder learning, as a complete death condition might not be necessary to explore to see the decreasing loss gradient leading to it. However, seeing a death condition could result in more extreme Q-predictions and thus more incentive to stay clear of a death.

7.8 Active control elements

There is ongoing research into adding active control elements to windturbine blades such as flaps. A difficulty incurred in that research is the design of a controller, as the high-dimensional action space

makes it difficult for a human to engineer a controller. As soon as we could provide a working version of an automatically learning, stable agent, we could assist the efforts of the people working on flap design to automatically learn controllers for them.

8 Conclusion

We showed that it is not trivially possible to extend existing continuous control reinforcement learning algorithms to provide windturbine control. However, we were able to reach near sensible control in two simplistic control tasks. To achieve this, we needed to modify existing algorithms to the pelicularities of windturbine control. Though our main task is not solved yet, we provided a promising strategy to handle death conditions in DDPG, which can alleviate the algorithm to scenarios in which the environment might step out of sensible operational ranges. We integrated PER into DDPG and at least gave a theoretical explanation why it works. We showed how to integrate QBlade into a reinforcement scenario. Though our results can not outperform an industry controller, we made some progress towards solving reinforcement learning on wind turbines.

Bibliography

[Aga] Abien Fred Agarap. Deep Learning using Rectified Linear Units (ReLU).

[BCP+] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym.

[Bel] Richard Bellman. The theory of dynamic programming. 60(6):503–516.

[Bis] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press ; Oxford University Press.

[BJSB] Tony Burton, Nick Jenkins, David Sharpe, and Ervin Bossanyi. *Wind Energy Handbook: Burton/Wind Energy Handbook*. John Wiley & Sons, Ltd.

[BPS+] Shalabh Bhatnagar, Doina Precup, David Silver, Richard S Sutton, Hamid R Maei, and Csaba SzepesvÃ¡ri. Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation. page 9.

[BW] J. J. Barradas Berglind and Rafael Wisniewski. Fatigue Estimation Methods Comparison for Wind Turbine Control.

[CKH+] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying Generalization in Reinforcement Learning.

[DWS] Thomas Degris, Martha White, and Richard S. Sutton. Off-Policy Actor-Critic.

[FGFGG] Borja Fernandez-Gauna, Unai Fernandez-Gamiz, and Manuel Grasa. Variable speed wind turbine controller adaptation by reinforcement learning. 24(1):27–39.

[FMP] Scott Fujimoto, David Meger, and Doina Precup. Off-Policy Deep Reinforcement Learning without Exploration.

[FvHM] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing Function Approximation Error in Actor-Critic Methods.

[HLD] Michael F. Howland, Sanjiva K. Lele, and John O. Dabiri. Wind farm power optimization through wake steering. 116(29):14495–14500.

[HMvH+] Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining Improvements in Deep Reinforcement Learning.

[Hub] Peter J. Huber. Robust estimation of a location parameter. 35(1):73–101.

[HZ] Yuenan Hou and Yi Zhang. Improving DDPG via Prioritized Experience Replay. page 10.

[JBMS] J. Jonkman, S. Butterfield, W. Musial, and G. Scott. Definition of a 5-MW Reference Wind Turbine for Offshore System Development.

[KB] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization.

[KJT] J. Z. Kolter, Z. Jackowski, and R. Tedrake. Design, analysis, and learning control of a fully actuated micro wind turbine. pages 2256–2263. IEEE.

[LHP+] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. CONTINUOUS CONTROL WITH DEEP REIN-

FORCEMENT LEARNING.

[MB] Michael C. Mozer and Jonathan Bachrach. Discovering the Structure of a Reactive Environment by Exploration. 2(4):447–457.

[MKS+a] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning.

[MKS+b] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. 518(7540):529–533.

[MWP+] D Marten, J Wendler, G Pechlivanoglou, C N Nayeri, and C O Paschereit. QBLADE: AN OPEN SOURCE TOOL FOR DESIGN AND SIMULATION OF HORIZONTAL AND VERTICAL AXIS WIND TURBINES. 3(3):6.

[PHD+] Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter Space Noise for Exploration.

[SAZFG+] Aitor Saenz-Aguirre, Ekaitz Zulueta, Unai Fernandez-Gamiz, Javier Lozano, and Jose Lopez-Guede. Artificial Neural Network Based Reinforcement Learning for Wind Turbine Yaw Control. 12(3):436.

[SB] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, second edition edition.

[SHM+] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. 529(7587):484–489.

[SLH+] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic Policy Gradient Algorithms. page 9.

[SLM+] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust Region Policy Optimization.

[SMSM] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. page 7.

[SQAS] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay.

[UO] G. E. Uhlenbeck and L. S. Ornstein. On the Theory of the Brownian Motion. 36(5):823–841.

[vHGH+] Hado van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude.

[vKPN+] G. A. M. van Kuik, J. Peinke, R. Nijssen, D. Lekou, J. Mann, J. N. SA¸rensen, C. Ferreira, J. W. van Wingerden, D. Schlipf, P. Gebraad, H. Polinder, A. Abrahamsen, G. J. W. van Bussel, J. D. SA¸rensen, P. Tavner, C. L. Bottasso, M. Muskulus, D. Matha, H. J. Lindeboom, S. Degraer, O. Kramer, S. Lehnhoff, M. Sonnenschein, P. E. SA¸rensen, R. W. Konneke, P. E. Morthorst, and K. Skytte. Long-term research challenges in wind energy“ a research agenda by the European Academy of Wind Energy. 1(1):1–39.

[Wil] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist rein-

forcement learning. page 28.

[ZLZH] Daochen Zha, Kwei-Herng Lai, Kaixiong Zhou, and Xia Hu. Experience Replay Optimization. page 7.

Acknowledgments

At first thanks to my advisors Matthew Lennie and Dmitrij Shlezinger for all the support they gave in the process of writing this thesis. Thanks to David Marten for helping with QBlade. Also, I guess, thanks David Silver - 7 of my most important papers were from you. Thanks to my laptop for almost dying, but then surviving. And big thanks to the mighty god of windturbines for letting me crush several thousand windturbines for science.