Sailing the Wind: Evaluating the Impact of COMA on Multi-Agent Active Wake Control in Wind Farms

What is the effect of COMA on the problem of AWC compared to single-agent RL algorithms?

Mihai Filimon

Supervisor(s): Mathijs de Weerdt, Grigory Neustroev

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Mihai Filimon
Final project course: CSE3000 Research Project
Thesis committee: Mathijs de Weerdt, Grigory Neustroev, Przemysław Pawełczak

Abstract

The close proximity of wind turbines to one another in a wind farm can lead to inefficiency in terms of power production due to wake effects. One technique to mitigate the losses is to veer from their individual optimal direction. As such, the wakes can be steered away from downstream turbines in order to increase the overall power output. Multi-Agent Reinforcement Learning (MARL) models the interactions between wind turbines and determines an optimal control strategy through agents that learn the collective consequences of their actions. To analyse the benefit of multi-agent cooperation and centralised critic evaluation, I investigated the effect of Counterfactual Multi-Agent Policy Gradients (COMA) on Active Wake Control. Ultimately, experiments on wind farms of three and sixteen turbines indicate that the algorithm performs moderately, yet worse than single-agent Reinforcement Learning. In addition, high computation costs hinder its application on real-life environments.

1 Introduction

In an effort to mitigate climate change, governments are increasingly shifting towards renewable energy. Especially in the following transition years, wind plays a crucial role in terms of energy production [1].

As wind farms grow in size, it is important to optimise energy production. However, wind farms face a significant issue – when wind turbines are placed directly in line after each other, their total power output is decreased. This is due to wake effects – areas of high turbulence and lower wind speed – caused by the extraction of wind by the first turbine [2] [3]. Managing and mitigating the negative consequences of wake effects can generate increased energy efficiency and improved renewable energy production. That, in turn, plays a pivotal role in reducing the dependency on fossil fuels, thereby contributing to combating climate change. Additionally, the current inefficiency caused by wake effects leads to decreased revenues that could otherwise be allocated to create a more sustainable society.

These implications can be tackled through an Active Wake Control (AWC) method - rotating the turbine in the horizontal plane in order to redirect the wake away from downstream turbines [4]. In this way, while the power output of the first turbine is lower, the total output is increased. Even though it would be also possible to take a passive approach and adapt the turbine and wind farm design based on specific local information [5], this research concentrates on the active approach of altering the direction of the wake.

To determine how and how much to adapt the orientation of the wind turbines in order to maximise the total output, we can turn to Reinforcement Learning (RL), and more specifically Multi-Agent Reinforcement Learning (MARL). Reinforcement Learning focuses on agents interacting with the environment and learning to make decisions through the rewards and penalties received. MARL, unlike single-agent RL, involves multiple agents that observe the collective environment. While single-agent RL has already been previously applied on wind farms in various contexts with remarkable achievements [2] [6] [7], it also faces a set of limitations. The training process needed to learn effective control policies can be time-consuming and computationally demanding, making it challenging to apply RL to real-time or large-scale AWC problems. [8].

Multi-Agent Reinforcement Learning (MARL) has the potential to address some of the limitations for AWC of single-agent RL. MARL can leverage the knowledge and expertise of multiple agents to improve the accuracy of models [9]. It enables agents to learn from each other's experiences by sharing their policies and observations, thus accelerating the process.

Out of the numerous existing MARL algorithms, I opted to investigate the Counterfactual Multi-Agent Policy Gradients (COMA) because of its ability to use a centralised critic along with distributed execution [10]. This feature allows agents to share information about the global state, which can lead to more effective coordination of their behaviour. Although using a centralised critic may increase the complexity of the algorithm, J. Foerster argues that COMA has a performance that is comparable to other cooperative algorithms [10]. As such, it is relevant to analyse the trade-off between centralised and decentralised critics, the limits of a centralised critic in terms of the number of agents, and whether it actually provides a relevant improvement compared to single-agent Reinforcement Learning.

Therefore, this paper focuses on the research question *"What is the effect of COMA on the problem of AWC compared to single-agent RL algorithms?"* with the following subquestions:

1. What is the difference in performance between COMA and TD3?
2. What are the limitations of COMA?

The remainder of this paper is organised as follows. Section 2 provides the background information on both Single-Agent and Multi-Agent Reinforcement Learning. Section 3 dives into the problem of AWC, while Section 4 presents the COMA algorithm in detail. The experiments are displayed in Section 5 and the results in Section 6. After that, Section 7 highlights some discussion points on responsibility and ethics. Lastly, the conclusions are presented in Section 8.

2 Background

2.1 Reinforcement Learning

Reinforcement learning (RL) is a branch of Machine Learning concerned with decision-making and control. Unlike supervised learning, where agents learn from labelled data, or unsupervised learning, which focuses on extracting patterns from unlabelled data, RL emphasises learning through interaction with an environment and receiving rewards as feedback [11]. This feedback guides the agent's behaviour towards maximising long-term rewards.

Reinforcement Learning is an evolving field, with ongoing research addressing various challenges. Relevant algorithms

such as Q-learning, Proximal Policy Optimisation (PPO) and Deep Deterministic Policy Gradient (DDPG) have demonstrated noteworthy achievements in game-playing or robotics [12].

Another state-of-the-art algorithm is TD3 (Twin Delayed DDPG), which Grigory Neustroev et al. use to tackle Active Wake Control through single-agent Deep RL [2]. It is an extension of DDPG with improved stability and performance, that addresses the overestimation of Q-values [13].

2.2 Multi-Agent Reinforcement Learning

Multi-Agent Reinforcement Learning (MARL) is an extension of RL that focuses on learning optimal policies for multiple interacting agents. In MARL, agents learn to make decisions based on their interactions with the environment and other agents. In general, MARL can be classified into three main approaches: fully cooperative, fully competitive and a combination of the two. Cooperative MARL maximises collective performance by promoting collaboration among agents, while in Competitive MARL agents aim to outperform each other [14].

Sharing experiences is a valuable mechanism of Multi-Agent Reinforcement Learning that can enhance the learning process and improve the performance of agents tackling similar tasks. This can be achieved through various means, such as communication between agents or imitation learning between observers and more proficient agents [14]. Besides, leveraging the decentralised structure of the task, parallel computation can significantly increase the pace of learning [14]. Moreover, MARL is inherently robust and facilitates the seamless insertion of new agents, which allows for high scalability and adaptability [14].

3 Active Wake Control

Wind turbines' proximity to one another in a wind farm affects the power production and mechanical stress due to the high turbulence created by the wake effects. Currently, the common practice is for each turbine to maximise its own power capture without taking into consideration the impact on neighbouring turbines [4]. However, this approach is not optimal for the total power production of the wind farm. In recent years, researchers have been developing a cooperative approach called Active Wake Control (AWC) to maximise the power production of the entire wind farm while reducing fatigue loading on the turbines.

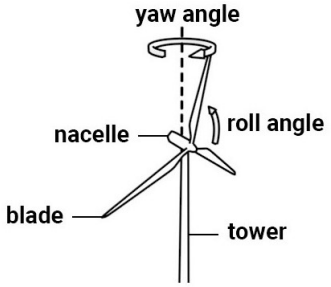

Figure 1: Turbine nomenclature designed by Ion Plămădeală.

There are two main types of AWC techniques. The first type intends to reduce the wake effect downstream by "reducing the axial induction factor of the upstream turbines, known as axial induction control or pitch-based AWC" [4]. This is achieved by increasing the blade pitch angle of the turbines on the windward side. Initial experiments have shown promising results in terms of increased power production, but recent simulations and tests have not confirmed these findings conclusively [4]. The second type of AWC, also known as yaw-based AWC, focuses on redirecting the wakes away from downstream turbines, which can be achieved by adjusting the yaw misalignment of the wind turbines. This research project aims to improve the latter.

4 COMA

Counterfactual Multi-Agent Policy Gradients (COMA) is an actor-critic Multi-Agent Reinforcement Learning (MARL) algorithm that uses a counterfactual baseline. COMA compares for each agent the global reward for the joint action with a counterfactual baseline, that "marginalises out a single agent's action while keeping the other agents' actions fixed" [10]. This way it learns more effectively and rapidly the real impact of different actions, and optimises the strategy [10].

While it would be simpler for each agent to be completely independent, by having its own actor and critic that consider only the agent's own action-observation history, COMA makes use of a centralised critic [10]. This decision encourages communication and coordination. As such, it allows the agents to deal with complex scenarios that require cooperation, and to understand the global consequences of their actions. In fact, the centralised critic is employed solely during the learning process, while during execution only the actor is required. The algorithm's architecture and information flow are illustrated in Figure 2.

COMA computes using the following formula an advantage function A on the joint action $\boldsymbol{u}$ and central state s for each agent a. On one hand, the critic adapts based on the global state, that encompasses the joint action-observation histories. On the other hand, while sharing parameters, each actor $\pi(u^a|\tau^a)$ is conditioned on its own action-observation history τ^a. COMA learns a centralised critic $Q(s, \boldsymbol{u})$ and compares the Q-value of the current action u^a to a counterfactual baseline [10].

$$A^a(s, \boldsymbol{u}) = Q(s, \boldsymbol{u}) - \Sigma_{u'^a} \pi^a(u'^a|\tau^a) Q(s, (\boldsymbol{u}^{-a}, u'a)) \quad (1)$$

To address the concerns of a high network size and expensive evaluations of the critic, COMA includes the actions of other agents as input to the neural network. Hence, "the counterfactual advantage can be calculated efficiently by a single forward pass of the actor and critic, for each agent" [10].

Therefore, given that agents in the Active Wake Control problem have interdependent objectives, this centralised approach can be particularly beneficial. However, that is not limited to AWC but, by design, COMA can be a good application in cooperative settings in general. By contrast, it would not be a suitable option for competitive scenarios.

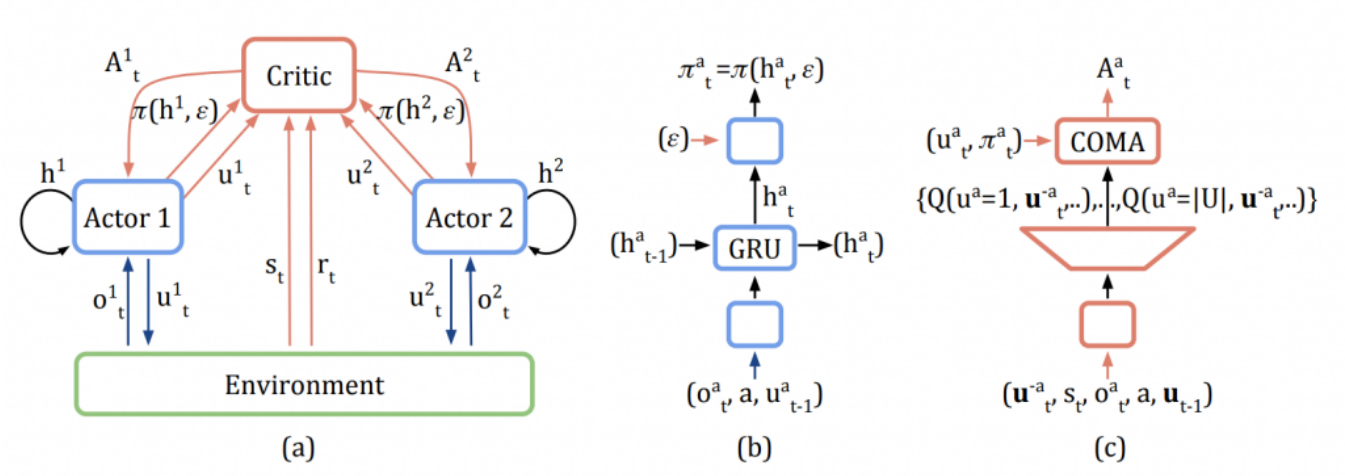

Figure 2: In (a), information flow between the decentralised actors, the environment and the centralised critic in COMA; red arrows and components are only required during centralised learning. In (b) and (c), architectures of the actor and critic [10].

5 Experimental Setup

Based on the explanations in the previous chapter, the hypothesis was that COMA can be applied to the Active Wake Control (AWC) problem through the wind farm environment developed by Grigory Nustroev et al. [15], and that it performs better than single-agent Reinforcement Learning and the AWC baseline. Nonetheless, it was expected that it would face difficulties as the wind farms increase in size, due to computational and communication costs of the centralised critic.

To analyse the performance, COMA was compared to the current state-of-the-art AWC as a baseline, which signifies turbines facing the wind directly, to TD3 as a single-agent RL algorithm, and to FLORIS – a wake-modelling framework with high accuracy in the optimisation of the wind farm layout by virtue of having complete information over the system [2] [13].

Hence, the prediction was that the algorithm performs for three turbines nearly as well as FLORIS, and that for 16 turbines it performs better than the baseline but worse than FLORIS and TD3.

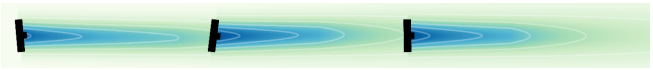

Figure 3: Visualisation of a "wind tunnel" of three turbines through the wind farm environment developed by Grigory Nustroev et al. [15].

In order to answer the research questions and test the hypotheses mentioned above, experiments were pursued in two different contexts: a basic one with only three turbines and a four-by-four grid of 16 turbines. The first scenario of three turbines in a row, also called a "wind tunnel" setup, was chosen aiming to model the extreme adversarial conditions [16]. The second one represents a larger wind farm, intended to be a threshold in terms of the capacity to handle complex situations and scalability.

All experiments were run using 5000 episodes with a maximum of 100 steps each. So, the agents (i.e. the wind turbines) take up to a hundred actions per learning experience. After that set of different changes in yaw, each agent learns their impact through the centralised critic, and this process is repeated 5000 times. It is important to mention that this implementation of COMA makes use of a discount factor value (gamma) of 0.99, which implies that future rewards are given higher importance compared to immediate rewards. Moreover, such a high value provides stability to the learning process, as the agent becomes less sensitive to variations in the immediate rewards.

6 Results

As can be seen from Figure 4, the agents are able to learn and increase the reward, i.e. the energy output, on the "wind tunnel" of three turbines. It is plausible that an even higher discount factor value could have led to even less sensitivity to noise in rewards, however, that possibility was not explored. COMA performs better than the AWC baseline, as it determines a policy that produces a higher energy output, however, the growth is still unsatisfactory in comparison with TD3. Therefore, in contrast to the hypothesis, the multi-agent cooperation with a centralised critic did not provide any significant improvement from single-agent Reinforcement Learning.

Similarly, Figure 5 shows that the 16 turbines arranged in a 4-by-4 grid perform modestly. As predicted in the beginning, COMA was capable of learning a policy that increases the energy output compared to the baseline, however, the results show that the improvement is far from the desired target of FLORIS and that TD3 still outperforms COMA. As such, since it did not manage to learn efficiently on a wind farm of 16 turbines, it is clear that COMA is unfit for large-scale wind farms.

The experiments demonstrated a series of limitations that COMA faces. Even with only 5000 episodes of 100 steps each, the algorithm takes a large amount of time to perform all the operations. The "wind tunnel" of 3 turbines requires

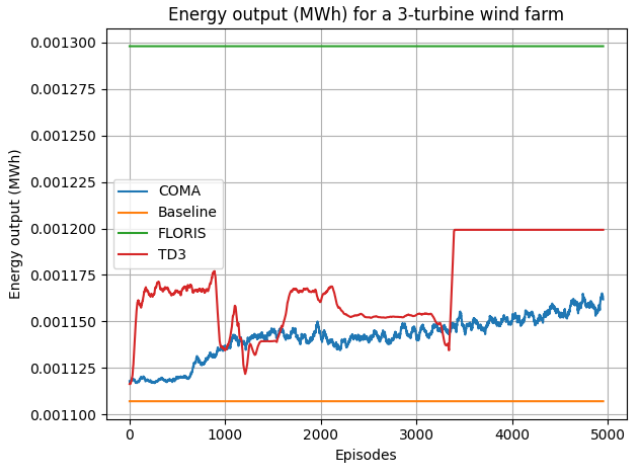

Figure 4: Learning curve of the energy produced by a "wind tunnel" of 3 turbines visualised with a moving average of window size 50.

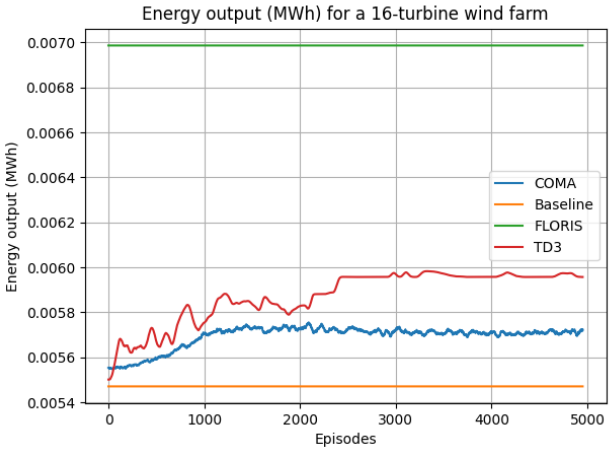

Figure 5: Learning curve of the energy produced by a 4-by-4 grid of 16 turbines visualised with a moving average of window size 50.

on average more than eight hours to finish, while the 4-by-4 grid needs almost 32 hours, which shows how computationally expensive COMA is. Consequently, there was not enough time to run the TD3 experiments, which is why the TD3 data presented in Figures 4 and 5 was used with the permission of Jasper van Selm. During this research, it was attempted to also test COMA on the "Princess Amalia" wind farm [17]. However, that was abandoned due to the substantial duration of the experiment, caused by additional personal technical constraints - running the algorithm on the Central Processing Unit (CPU) - which is considerably slower compared to a Graphics Processing Unit (GPU). Hence, based on COMA's behaviour on two simulation environments of limited size, its application in large environments would require days to complete even a short training period. It does not provide the necessary scalability for real-world wind farms, confirming the initial hypothesis.

7 Responsible Research

The COMA algorithm is available in multiple variants on GitHub. While the original version was published by Oxford University, the most compatible implementation for the AWC problem, which was used during this research project, was developed by Matteo Karl Donati [18]. The wind farm environment created by Grigory Neustroev et al. is also openly accessible through the repository of Delft University of Technology [15]. As such, anyone can build upon their algorithm in a similar manner to the one I pursued. Furthermore, the placement of the "Princess Amalia" wind farm is available through the "Hollandse Kust Nord B" data set of the Netherlands Enterprise Agency (RVO) [17]. As all the components are open source, they are "findable, accessible, interoperable and reusable" (FAIR) [19]. Therefore, the research is completely reproducible and verifiable by any individual.

From an ethical point of view, the research does not raise any significant concerns in terms of societal prejudices. The data revolves around inanimate objects that have no direct interaction with humans. Hence, the eventual biases are solely technical. However, as the data was published by a governmental agency, it is important to mention that, out of time constraints on the research project, it was not possible to assess whether the data represents the real case situation.

8 Conclusion and Future work

Active Wake Control remains an important challenge to be solved and a promising real-life application for Multi-Agent Reinforcement Learning (MARL) algorithms. Nonetheless, COMA's centralised-critic approach combined with counterfactual comparisons did not perform well in any wind farm environment. The algorithm poses significant computation constraints that do not guarantee an advancement from single-agent Reinforcement Learning.

Still, the experiments have been limited to uncomplicated settings and have not included real-life scenarios, such as wind coming from different directions or the placements of already existing wind farms. Therefore, to properly determine whether MARL is a solution for Active Wake Control, it is vital for future researchers to continue investigating the potential of COMA by applying it on a real wind farm environment, such as "Princess Amalia" or "Gemini".

Besides the fact that there was no investigation of real-life settings, the research was pursued as part of a bachelor's degree thesis. Taking into account the lack of prior knowledge and experience, it is recommended that the research is reproduced and that COMA shall be researched further. While the experiments provided relevant information, it is not possible to draw a complete conclusion yet regarding COMA's potential for Active Wake Control.

ACKNOWLEDGEMENTS

I would like to thank Professor Mathijs de Weerdt for the feedback on this paper. I am grateful to Grigory Neustroev for his guidance and support in the research project, to Ion and Jasper for openly sharing the turbine design and TD3 outcomes, and to everyone in the team for continuously sharing ideas and pushing each other.

References

[1] S. Potrč, L. Čuček, M. Martin, and Z. Kravanja, "Sustainable renewable energy supply networks optimization–the gradual transition to a renewable energy system within the european union by 2050," *Renewable and Sustainable Energy Reviews*, vol. 146, p. 111186, 2021.

[2] G. Neustroev, S. P. Andringa, R. A. Verzijlbergh, and M. M. de Weerdt, "Deep reinforcement learning for active wake control," *International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*, vol. 322, no. 10, pp. 944–953, 2022.

[3] L. Vermeer, J. Sorensen, and A. Crespo, "Wind turbine wake aerodynamics," *Progress in Aerospace Sciences*, pp. 467–510, 2003.

[4] S. Kanev, F. Savenije, and W. Engels, "Active wake control: An approach to optimize the lifetime operation of wind farms," *Wind Energy*, vol. 21, no. 7, pp. 488–501, 2018.

[5] M. F. Howland, S. K. Lele, and J. O. Dabiri, "Wind farm power optimization through wake steering," *Proceedings of the National Academy of Sciences*, vol. 116, no. 29, pp. 14 495–14 500, 2019.

[6] H. Dong, J. Zhang, and X. Zhao, "Intelligent wind farm control via deep reinforcement learning and high-fidelity simulations," *Applied Energy*, vol. 292, pp. 2974–2982, 2021.

[7] P. Stanfel, K. Johnson, C. J. Bay, and J. King, "Proof-of-concept of a reinforcement learning framework for wind farm energy capture maximisation in time-varying wind," *Journal of Renewable and Sustainable Energy*, p. 13:043305, 2017.

[8] R. Nian, J. Liu, and B. Huang, "A review on reinforcement learning: Introduction and applications in industrial process control," *Computers & Chemical Engineering*, vol. 139, p. 106886, 2020.

[9] L. Buşoniu, R. Babuška, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[10] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," *AAAI Conference on Artificial Intelligence, AAAI-18*, p. S0306261921004086, 2017.

[11] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] S. Dankwa and W. Zheng, "Twin-delayed ddpg: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent," in *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, ser. ICVISP 2019. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3387168.3387199

[14] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

[15] G. Neustroev, S. P. Andringa, R. A. Verzijlbergh, and M. M. de Weerdt, "wind-farm-env," 2022. [Online]. Available: https://github.com/AlgTUDelft/wind-farm-env

[16] L. Machielse, "Controlling wind, tunnel theorie," ECN-E-11-065, Energy Research Center of the Netherlands, ECN, Tech. Rep., 2011.

[17] RVO, "Hollandse kust noord (site b) dataset," 2019. [Online]. Available: https://offshorewind.rvo.nl/files/view/00419133-21ab-470b-bfe6-c010dfc76c66/1567080325hkn_20190815_fugro_hknb_processed%20data.zip

[18] M. K. Donati, "Pytorch implementation of counterfactual multi agent policy gradients," 2020. [Online]. Available: https://github.com/matteokarldonati/Counterfactual-Multi-Agent-Policy-Gradients

[19] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.