Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Cost Efficient Distributed Load Frequency Control in Power Systems

Flavio R. de A. F. Mello, Dimitra Apostolopoulou, Eduardo Alonso

City, University of London
London, UK EC1V 0HB
Email: {flavio.ribeiro-de-aquino-f-mello, dimitra.apostolopoulou, e.alonso}@city.ac.uk

Abstract: The introduction of new technologies and increased penetration of renewable resources is altering the power distribution landscape which now includes a larger number of micro-generators. The centralized strategies currently employed for performing frequency control in a cost efficient way need to be revisited and decentralized to conform with the increase of distributed generation in the grid. In this paper, the use of Multi-Agent and Multi-Objective Reinforcement Learning techniques to train models to perform cost efficient frequency control through decentralized decision making is proposed. More specifically, we cast the frequency control problem as a Markov Decision Process and propose the use of reward composition and action composition multi-objective techniques and compare the results between the two. Reward composition is achieved by increasing the dimensionality of the reward function, while action composition is achieved through linear combination of actions produced by multiple single objective models. The proposed framework is validated through comparing the observed dynamics with the acceptable limits enforced in the industry and the cost optimal setups.

Keywords: Multi-Agent Reinforcement Learning, Multi-Objective Reinforcement Learning, Frequency Control, Economic Dispatch, Deep Deterministic Policy Gradient

1. INTRODUCTION

Over recent years, the field of electrical power systems has been experiencing the beginning of what may prove to be a structural transformation. Renewable resources have been increasing their penetration in the marketplace, which may displace traditional sources. Decreasing costs of solar panels lead to increased adoption in households, to the extent that there are already legal provisions for household customers to sell stored energy back into the electrical grid as mentioned in Ambrose (2019). Vehicle to grid and smart charging technologies are posed to enable electric cars to contribute to balancing the power grid see, e.g., Steitz (2019). This represents a significant increase in the complexity of the grid, shifting away from a small number of large scale producers to include an ever increasing number of micro-sized sources in the form of individual households, electric cars, etc. Such manifold structure, in turn, will intensify the need for intelligent, automated and decentralized control solutions. Modern electrical energy distribution is largely done by means of wide-ranging synchronous grids. Being synchronous means the entirety of the grid is electrically connected and thus every element attached to the grid shares the same observed operating frequency. This is true for both the consumers as well as the producers (generators). In these systems the observed operating frequency changes over time according to i) the total power being injected into the system by all the generators; ii) the total power being consumed by all loads. To electrically balance the system, independent system operators (ISOs) send signals to generators to modify their output such that load and generation are balanced and the system frequency is nominal.

Multiple techniques have already been proposed to achieve frequency control decentralization. From a traditional control standpoint, Apostolopoulou et al. (2015a) propose methods for approximating the automatic generation control (AGC) algorithm while solving the economic dispatch in semi-decentralized fashion by restricting the Balancing Authority (BA) areas communication and, thus, avoiding congestion associated with the exponential increase of connections in the network (see Apostolopoulou et al. (2015a) and Apostolopoulou et al. (2015b)). Additionally, Model Predictive Control (MPC) techniques have been proposed to perform decentralized frequency control whilst satisfying predetermined constraints (see Ali et al. (2017), Kumtepeli et al. (2016) and Heydari et al. (2019)). In the Reinforcement Learning realm, Rozada (2018) proposes the use of Multi-Agent Reinforcement Learning (MARL) techniques, more specifically, the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm, which proved able to successfully perform primary and secondary control but failed to perform tertiary control. Despite being separate layers of control, both primary and secondary control share a common overarching objective related to frequency deviation. Tertiary control, however, is associated with a slightly different objective: to minimize the total cost of electricity production. These differences in objective alignment could explain why the MAADPG algorithm, as implemented in said paper, successfully performed primary and secondary controls but failed with tertiary control. For this end, this paper proposes the addition of Multi-Objective Reinforcement Learning (MORL) techniques to the algorithm.

In this paper we propose an Reinforcement Learning (RL) technique for training autonomous, decentralized agents able to perform frequency control in an electric power system to: i) maintain the system frequency within predefined tolerated limits; ii) minimize the cost of production. More specifically, we frame the frequency control problem as a Markov Decision Process to allow for the use of reinforcement learning techniques (Section 2); ii) propose the incorporation of two distinct multi-objective reinforcement learning techniques to the MADDPG algorithm to perform frequency control in a cost-efficient way (Section 3); iii) compare the performance of both techniques through numerical studies (Section 4); and iv) draw conclusions from the observed behaviours (Section 5).

2. BACKGROUND

In this section, the frequency control problem is formulated and the reinforcement learning techniques employed to perform such control are presented.

2.1 Load frequency control and Economic Dispatch

Frequency control can be divided into three hierarchical layers: Primary, Secondary and Tertiary control. Primary control acts to counterbalance changes in the total system load by adjusting the output levels of all generators attached to the grid by an amount proportional to the difference between the observed and nominal frequency Miller and Malinowski (1994). Primary control has the benefit of being completely decentralized as each generator is able to observe individually the current frequency in the system. Secondary control or Automatic Generation Control (AGC) systems act upon the steady state error resulted by the limitations in primary control. These algorithms are often centralized to some extent, with an individual entity overseeing the entire grid and issuing commands for the individual generators. The power system dynamics of the secondary control system are as follows:

$$
\begin{gathered}
P(t+1) = P(t) + \frac{Z_{\text{total}}(t) - \frac{1}{R_D}\Delta\omega(t) - P(t)}{T_G}, \\
\omega(t+1) = \omega(t) + \frac{P(t+1) - L(t) - D\Delta\omega(t)}{M}, \\
Z_{\text{total}}(t) = \sum_{i=1}^{I} Z_i(t), L_{\text{total}}(t) = \sum_{j=1}^{J} L_j(t), \\
\Delta\omega(t) = \omega(t) - \omega_{\text{nominal}},
\end{gathered}
$$

where $P(t)$ is the total power injected into the grid at time t, $Z(t)$ is the secondary control action, $Z_i(t)$ is the control action of generator i at time t, $L(t)$ is the load at time t, $L_j(t)$ is the load by consumer j at time t, $\omega(t)$ is the system frequency at time t, R_D is the droop control coefficient selected for the system, D is the damping coefficient of the system, M is the electrical inertia of the grid, I is the number of generators, J is the number of loads in the system, and ω_{nominal} is the nominal frequency.

Also referred to as Economic Dispatch, the objective of tertiary control is to minimize the total production costs of the grid. Doing so requires a centralized entity with knowledge of each generating power output and cost of production curve, as well as relevant physical limits with regards to minimum and maximum output levels. For a power system with I generators and J loads, the following set of equations apply at time t:

$$
\begin{aligned}
&\min_{p_i(t)} C_{\text{total}}(t) = \sum_{i=1}^{I} C_i(t) = \sum_{i=1}^{I} \big(\alpha_i + \beta_i p_i(t) + \gamma_i p_i^2(t)\big), \\
&\text{s.t.} \sum_{i=1}^{I} p_i(t) = \sum_{j=1}^{J} L_j(t), \\
&\qquad p_i^{min} \le p_i(t) \le p_i^{max}, \text{ for } i = 1, \dots, I,
\end{aligned}
$$

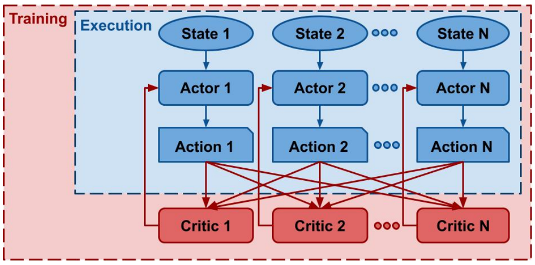

Fig. 1. Actor/Critic relationship in MADDPG

where $C_{\text{total}}(t)$ is the total cost of production at time t, $C_i(t)$ is the cost of production of generator i at time t, $p_i(t)$ is the output level of generator i at time t, α_i, β_i, γ_i are constants, p_i^{min} and p_i^{max} are, respectively, the minimum and maximum output levels for generator i, and $L_j(t)$ is the power consumption of load j at time t. Solving tertiary control entails finding the power output combination set $\{p_1^\star(t), p_2^\star(t), p_3^\star(t), \dots, p_I^\star(t)\}$ that minimizes the global cost $C_{\text{total}}(t)$ while respecting the constraints of keeping the system balanced and every generator output within its operational limits. The system reaching steady state operation entails that $\Delta\omega(t) = 0$ and, therefore, $P(t) = Z_{\text{total}}(t) = \sum_{i=1}^{I} p_i^\star(t)$.

2.2 Reinforcement Learning

Reinforcement Learning can be defined as a family of techniques used to train agents based on their interactions with the environment and the associated rewards/punishments observed. Given enough observations the trained agent becomes able to issue commands so as to find an optimal policy. The problem approached in this paper can be classified as fully cooperative as the agents work together to reach two objectives: i) electrically balance the system within the tolerated range indicated by the frequency deviation; ii) minimize the total cost of production. MORL relates to RL problems with multiple, sometimes conflicting, objectives. Successfully trained MORL agents should be able to perform tradeoffs, intentionally sacrificing adherence to one objective while advancing towards a more desired global state. To this end, there are a number of different techniques that can be employed, ranging from weighted-sum to Pareto dominating policies see, e.g., Liu et al. (2015), and Moffaert and Nowé (2014). The choice of which approach to take becomes an integral part of the design process of the solution.

The technique used in this paper is named MADDPG and is considered an extension of Deep Deterministic Policy Gradient (DDPG), combined with some elements of actor-critic RL techniques see, e.g., Lowe et al. (2017). The MADDPG algorithm applies the actor-critic concept to multi-agent scenarios by centralizing learning whilst decentralizing execution, see Fig. 1. Once trained, the agents rely solely on their actors to take actions in the execution environment. Actors, therefore, remain decentralized in nature, having access only to the same information said agent would have in execution time. The critics, however, are centralized and have additional information in the form of the actions taken by all the other actors in the system.

3. PROPOSED FRAMEWORK

In this section, the details of the proposed implementation are described. This includes the neural networks architectures, the guidelines used for determining the reward functions used, and the approaches taken to incorporate multi-objective capabilities in the trained agents.

3.1 Neural Networks

The MADDPG algorithm leverages fully connected deep neural networks to model both the actor and the critic. In this study, both networks follow the same schema, with slight changes in the input/output layers. Additionally, this study employs the same algorithm to learn different policies to achieve different objectives. Often this requires changes in both the reward function and the set of variables that compose the $S_i(t)$ input, i.e., the state observed by agent i at time t. These changes are further described in Section 4 on a case by case basis. Common among all case studies are the output layers. The actor network outputs the action, in the form of change in total secondary action $(\Delta Z_i(t))$, where $Z_i(t) = \Delta Z_i(t) + Z_i(t-1)$, to be taken by its respective generator at time t. The critic network takes as input the outputs from all actor networks $(\Delta Z_0(t), \ldots, \Delta Z_I(t))$ and outputs the estimated quality (Q-value) for that state-action for its respective generator. The base neural networks used are depicted in Fig. 2.

3.2 Reward Function Design

Reward functions play a pivotal role in the success of Reinforcement Learning models. In multi-objective scenarios, the proportion between each reward component has increased importance. With these characteristics in mind, a collection of guiding principles shaped the design process of the reward functions used, namely: Finite upper and lower bounds act as points of reference for comparing given rewards, facilitating the assessment of their quality. Define the global reward function as a composition, of individual reward functions designed for each objective. While keeping the adherence to all other objectives constant, increasing adherence to a given objective should monotonically increase the total reward. This is only possible if the objectives are not intrinsically contradictory. Having the global maxima of all individual objectives reward functions coincide means that the state which provides the maximum reward globally is the same which maximizes rewards for all individual objectives. For the purpose of streamlining the design process, all individual reward functions share the same base function $f(x) = a2^{-bx^2}$, where x is the input of the reward function, which varies according to the objective (e.g., $\Delta\omega$ for balancing frequency), and a and b are parameters in $\mathbb{R}^+$. This function provides some useful traits: It is symmetric with respect to the y-axis, which is instrumental if the objective is to minimize deviation. Besides, the base function has one single maximum at the origin, which means that composition by either multiplication or addition retains a single global maximum at the same point. Finally, parameters a and b can be used ad hoc for deforming the function while keeping the symmetry and maximum location characteristics.

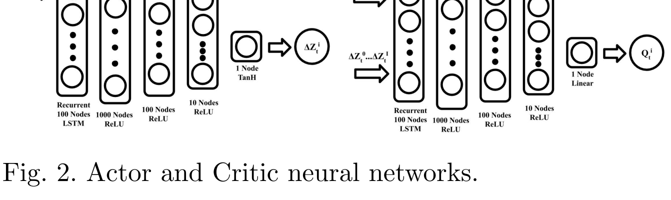

Fig. 2. Actor and Critic neural networks.

3.3 Multi-Objective

This investigation sets out to test two distinct strategies for obtaining multi-objective optimization: reward-composition and action-composition. The former strives to accomplish the overall objective by learning a single policy that is able to fulfil multiple objectives. This is achieved by consolidating multiple objectives and their hierarchical relationship into a single reward function. Conversely, the action-composition approach trains one single-purpose set of agents per objective. During execution, actions from all sets of agents are consolidated into individual final actions. For a system with K agents and M objectives this composition is expressed as:

$$A_k(t) = \sum_{m=1}^{M} \rho_m \tilde{A}_k^m(t), \sum_{m=1}^{M} \rho_m = 1, 0 \leq \rho_m \leq 1,$$

where $A_k(t)$ is the action to be taken by agent k at time t, ρ_m is the weight given to objective m, $\tilde{A}_k^m(t)$ is the action assigned to agent k, at time t, by the model aimed at optimizing objective m.

When performing action-composition, the reward functions used for each overarching objective does not intrinsically carry information regarding such preferences, these are declared in the form of the weights ρ_m, for $m = 1, ..., M$ used in runtime. One prerequisite for performing action composition is for the action-space to be quantitative. In categorical action environments, action consolidation cannot be done via arithmetic operations.

3.4 Reward Composition vs Action Composition

We propose two different methods of achieving multi-objective learning, reward composition and action composition. Besides observed performance, there are multiple factors that are taken into account when choosing a technique to be used in an industrial setting. In that sense, it can be argued that the action composition approach is superior from a systems design standpoint. Among the benefits provided by this strategy, one can single out the following: Breaking down the global model into a single objective ones decreases coupling between the models, facilitates reuse, and simplifies debugging (Separation of Concerns). Crafting bespoke multi-objective reward functions is a time-consuming enterprise. Breaking down into single objective rewards could speed up development as single objective reward functions behave in a more predictable way (Simplified Modeling). Declaring the objective priorities at the runtime means that these priorities can be seamlessly changed. Furthermore, finding the optimal priorities ratio can be done faster as the test feedback loop is tighter (Variable Priorities). Individual models can have different inputs. If different objectives of the system are associated with different Service Level Agreements (SLAs), the information sources which provide the inputs can be designed to match these SLAs. In a single model, all inputs are necessary to sample the actions, therefore have to provide an SLA that is compatible with the most critical objective. Using the studied scenario as an example, balancing the system frequency is critical at all times while optimizing for cost albeit still important is something that can be overlooked in critical situations. If those objectives are tackled by individual models, the inputs for balancing the system should be kept available and with minimum delay at all times. Conversely, the inputs for optimizing the cost can have their requirements relaxed — if they become offline, the system still can be operated at a degraded level by relying only on the frequency balancing model (Separate Data Sources).

3.5 Decentralization

We are using Multi-Objective RL techniques to solve primary, secondary, and tertiary control in a multi-agent-based model. The system designed in this analysis, albeit decentralized from the decision-making standpoint, still relies on some centralized information regarding the current state of the system, in particular $Z_{\text{total}}(t)$ the secondary control action at each time t. Although not completely fulfilling the decentralization requirement, this marks an important step towards full decentralization, as it changes the nature of the centralized entity from a fully-fledged decision maker to an information broker.

4. NUMERICAL STUDIES

The software developed for performing these case studies is fully configurable and allows for further experimentation with different configurations for electrical systems with any number of loads and generators electrical constants, and even reward functions and state inputs. The source code is open for future use and can be found at `https://github.com/melloflavio/2019-MSc_Thesis`.

4.1 Electrical System

We performed a multitude of experiments aimed at assessing the feasibility of leveraging multi-objective techniques to perform primary, secondary and tertiary control in an electrical power system. In order to perform the control experimentation, an electrical system simulator was implemented according to the equations described in Section 2.1. A consistent system topology was used across all experiments: three generators (G1, G2, and G3) and one single load (L1). The electrical constants were also kept the same for all the experiments; $R_D = 0.1$ pu, $T_G = 30$ s, $D = 0.016$ pu, and $M = 0.1$ pu. In this context, pu refers to the 100 MVA base power used throughout this paper.

Each simulation episode begins at $t = 0$ considering that the system is fully balanced ($P(0) = L(0)$, $\Delta\omega(0) = 0$) and a perturbation occurs at t_0 in the form of a change in the total load. The task being performed then is to balance the system after this initial perturbation. In the interest of increasing the robustness of the models trained, the application developed is able to introduce noise in the simulated environment in the form of changing the initial values for the loads and generators power levels. The noise takes the form of a uniform distribution with magnitude of 0.5% of the initial value. The models were trained by running simulations lasting 15000 episodes each.

For each generator, a distinct cost profile was selected with the purpose of ensuring that the optimal setup is such that no generator is in either minimum (0.5 pu) or maximum (3.0 pu) output values. Table 1 indicates the cost profiles of all generators:

Table 1. Generator Cost Profiles

Generator	α [\$/h]	β [\$/(h $\cdot$ MVA)]	γ [\$/(h $\cdot$ MVA^2)]
G1	510.0	7.7	0.00142
G2	310.0	7.85	0.00194
G3	78.0	7.55	0.00482

4.2 Case Study I - Frequency Control

In this study the objective was to minimize frequency deviation $\Delta\omega(t)$ at each time t which is the only state input. Based on Section 3.2 the reward function used was: $r_I(\Delta\omega(t)) = \left(9 \cdot 2^{-\frac{\Delta\omega^2(t)}{2}} + 2^{-\frac{\Delta\omega^2(t)}{100}}\right)\frac{1}{10}$. The results are depicted in Fig 3. After approximately 20 seconds, the load was successfully balanced and the power output and system frequency oscillates within 0.05 Hz (0.1%) of the nominal setpoint, which falls inside the accepted range of 0.5 Hz (1%) established by National Grid Electricity Transmission (2017).

In this case study we see that the two generators learn to reach their minimum output as fast as possible (Fig. 4), while the third generator controls its output to stabilize the system gradually reducing the steady-state error. This "cooperation by omission" approach does not appear to be the most efficient way to balance the system. One possible reason for this behaviour could be that reaching the maximum/minimum limits may be the best way to ensure stable output for the other generators, as these limits are enforced in the simulation, and not in the modelled neural networks themselves (i.e., once the secondary action reaches whichever limit, the neural network may still issue commands to go beyond such limits, but they are disregarded by the electrical system simulation). Future work will focus on training with more diverse loads that better cover the full spectrum of the systems total power capacity to obtain more robust cooperative strategies.

4.3 Case Study II - Reward Composition: Cost and frequency deviation minimization

This case study follows the reward-composition strategy where the state used as input in the algorithm is a triplet containing $\Delta\omega(t)$, $Z_i(t)$ and $Z_{\text{total}}(t)$. Additionally, a single reward function that reflects both objectives was crafted following the guidelines set in Section 3.2 and may be written as follows:

$$r_{II}(\Delta P_{\text{total}}(t), \Delta\omega(t)) = f(\Delta P_{\text{total}}(t))g(\Delta\omega(t)), \qquad (1)$$

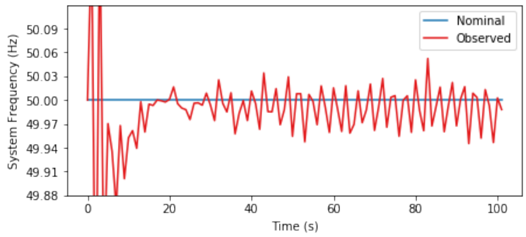

Fig. 3. Case I: Observed frequency

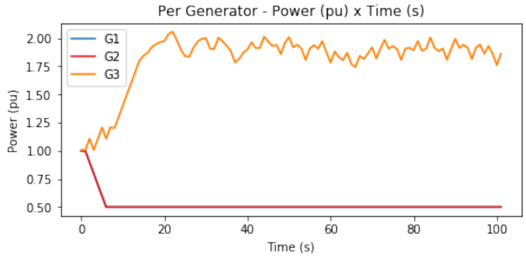

Fig. 4. Case I: Generator output

$$f(\Delta P_{\text{total}}(t)) = 2^{\frac{-\Delta P^2_{\text{total}}(t)}{4}}, \qquad (2)$$

$$g(\Delta\omega(t)) = \left(9 \cdot 2^{-\frac{\Delta\omega^2(t)}{2}} + 2^{-\frac{\Delta\omega^2(t)}{100}}\right)\frac{1}{10}. \qquad (3)$$

The frequency component — $g(\Delta\omega(t))$ — is similar to the function used in case study I. The cost component — $f(\Delta P_{\text{total}}(t))$ — is expressed in terms of the total power deviation from the cost optimal setup with a normalization component denoted by $\Delta P_{\text{total}}(t) = \sum_{i=1}^{I} \left| \frac{p_i(t)}{p_i^\star(t)} - 1 \right|$, where $p_i(t)$ is the power produced by generator i at time t, and $p_i^\star(t)$ is the power output of generator i at time t which minimizes the total cost for the total output of all generators observed at time t.

The frequency, see Fig. 5, remains within 0.12 Hz the nominal value, and exhibits a consistent downward shift of approximately 0.05 Hz. This falls inside the the accepted range of 0.5 Hz (1%) established by National Grid Electricity Transmission (2017). As seen in Fig. 6, Generators G1 and G2 follow closely their optimal outputs for the given total output at any given point. While G3 moves directly to and remains at the minimum output. Such behaviour could be interpreted as being associated with G3's optimal output being close enough to the minimum value that the model as a whole benefits more by having G3 remain at a flat level, and thus providing more certainty to G1 and G2, than by actively attempting to follow its optimal value.

4.4 Case Study III - Action Composition: Cost and frequency deviation minimization

This case study tests the action-composition strategy where for each overarching objective, one set of agents is trained. Set 1, aimed at balancing the system load, is in fact the same model trained in case study I. In Set 2, the state input is composed by the duple $Z_i(t)$, $Z_{\text{total}}(t)$. The model is trained with a single objective reward function aimed at finding the minimum cost of production for every total output as seen below:

$$r_{III-2}(\Delta P_{\text{total}}(t)) = \left(9 \cdot 2^{-\frac{\Delta P^2_{\text{total}}(t)}{2}} + 2^{-\frac{\Delta P^2_{\text{total}}(t)}{100}}\right)\frac{1}{10}. \qquad (4)$$

In order to test the performance of the action composition approach we choose two scenarios: i) Frequency Dominant: $\rho_{frequency} = 0.7$, $\rho_{cost} = 0.3$; ii) Cost Dominant $\rho_{frequency} = 0.3$, $\rho_{cost} = 0.7$.

In the frequency dominant study, as in case study I, the trained frequency model relies basically on a single generator to provide most of the output and change its output to gradually balance the system. Furthermore, in this particular instance of the trained model, the generator elected for that role was G3, which also has the characteristic of being the least cost-efficient generator among the set. Together, these characteristics result in a clashing behaviour between both models (Figs. 7, 8). For generators G1 and G2, the frequency model simply acts to reduce the power indefinitely, relying on the enforcement of the minimum floor. The mixing weights are such that the frequency model continuously overrides the actions issued by the cost model. G3 initially rises much like in the frequency model. As it approximates the output which would balance the system, the frequency model issues increasingly smaller actions to perform the fine-grained balance of the system. Meanwhile, the cost model continues to issue actions to dramatically lower G3's by virtue of it being the least cost-effective generator and having an output significantly above its optimal value. These divergent actions eventually reach an equilibrium at a point in which the frequency is far enough from the nominal so that the magnitude of the frequency and cost actions are counterbalanced.

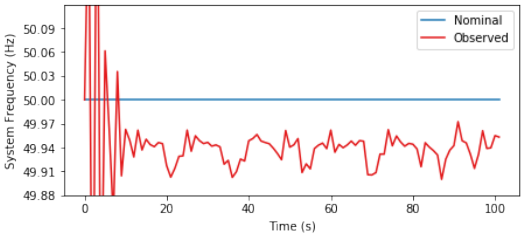

Fig. 5. Case II: Observed frequency

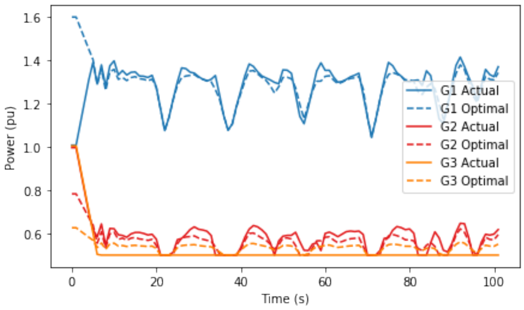

Fig. 6. Case II - Generator output vs cost optimal

In the cost dominant study, the system is balanced within 0.03 Hz of the nominal setpoint, while the power output levels approach those that lead to the minimum cost of production (see Figs.9, 10). In this case, the downward shift in frequency seen in the frequency dominant test is no longer observed.

5. CONCLUSIONS

In this paper, we formulated the load frequency control problem as a Markov Decision Process and employed reinforcement learning techniques to train autonomous agents able to perform decentralized primary, secondary and tertiary control. We then proposed two strategies for dealing with the tradeoffs associated with multiple objectives, each with its own benefits and disadvantages. Reward Composition consolidates multiple objectives into a single reward function used to train a single set of models, whereas Action Composition trains one set of models per objective and then consolidates the actions issued by all sets. Both methodologies decentralize decision making, but retain some degree of centralization in the

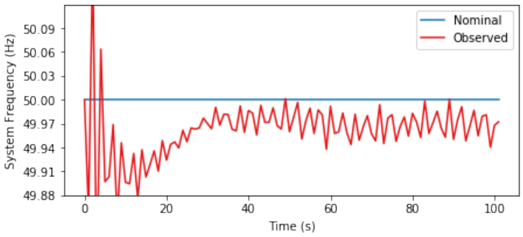

Fig. 7. Case Study III (Frequency Dominant): Observed frequency

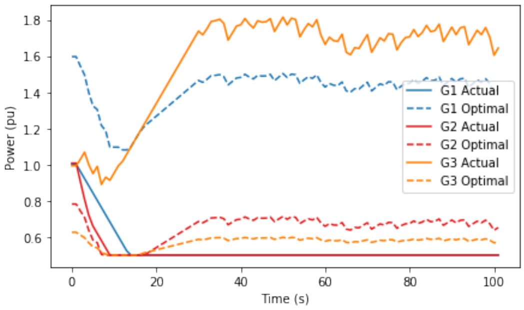

Fig. 8. Case Study III (Frequency Dominant): Generator output vs cost optimal

form of the total secondary action used in the state input for the models during the training phase. Overall both approaches were able to restore the system frequency in a cost efficient way.

More work would be required to assess the reproducibility of such results in high-fidelity simulations, demonstrating its generalizable capabilities, applying it to industrial scenarios and comparing it with existing control solutions employed in the industry. Furthermore increasing the number of generators significantly increased the computational power required for training. Further work could deal with such scalability issues at the model/algorithm level (e.g., increase sample efficiency, or a training curriculum), as well as a system architecture one (e.g., deploy replicas of pretrained models). The algorithm of choice, MADDPG, intensifies the scalability issues by requiring a single critic to be trained for every actor, rather than relying on a global critic for all actors. Moreover, being it an online algorithm, when deploying in real-world applications, one would encounter sub-optimal performance during training. To circumvent that, one could pretrain models in highly detailed simulation environments and deploy to real-world applications once the models achieve sufficiently consistent and acceptable performance. Additionally, one could also perform tests with different neural network architectures to assess its impact in the observed performance. Future research includes the introduction of more objectives, such as ecological impact of powering the grid, as the methodology employed and codebase developed have no restriction regarding the number of objectives being pursued. Regarding decentralization, one possibility would involve the use of accessory metadata such as timestamps associated with the total secondary action. Intuitively, this could help relax the real-time constraint of the information centralization by enabling agents to rely on offline information.

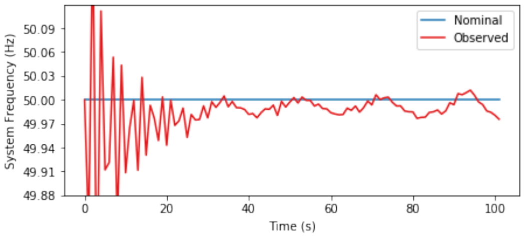

Fig. 9. Case Study III (Cost Dominant): Observed frequency

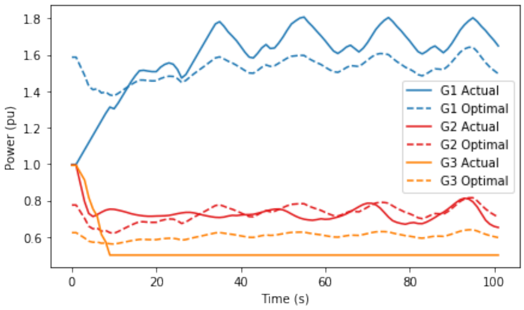

Fig. 10. Case Study III (Cost Dominant): Generator output vs cost optimal

REFERENCES

Ali, A., Khan, B., Mehmood, C.A., Ullah, Z., Ali, S.M., and Ullah, R. (2017). Decentralized mpc based frequency control for smart grid. In *2017 International Conference on Energy Conservation and Efficiency (ICECE)*, 1–6. doi:10.1109/ECE.2017.8248819.

Ambrose, J. (2019). New rules give households right to sell solar power back to energy firms. URL https://www.theguardian.com/environment/2019/jun/09/energy-firms-buy-electricity-from-household-rooftop-solar-panels.

Apostolopoulou, D., Sauer, P.W., and Domnguez-Garca, A.D. (2015a). Balancing authority area coordination with limited exchange of information. In *2015 IEEE Power Energy Society General Meeting*, 1–5. doi:10.1109/PESGM.2015.7286133.

Apostolopoulou, D., Sauer, P.W., and Domnguez-Garca, A.D. (2015b). Distributed optimal load frequency control and balancing authority area coordination. In *2015 North American Power Symposium (NAPS)*, 1–5. doi:10.1109/NAPS.2015.7335113.

Heydari, R., Khayat, Y., Naderi, M., Anvari-Moghaddam, A., Dragicevic, T., and Blaabjerg, F. (2019). A decentralized adaptive control method for frequency regulation and power sharing in autonomous microgrids. In *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, 2427–2432. doi:10.1109/ISIE.2019.8781102.

Kumtepeli, V., Wang, Y., and Tripathi, A. (2016). Multi-area model predictive load frequency control: A decentralized approach. In *2016 Asian Conference on Energy, Power and Transportation Electrification (ACEPT)*, 1–5. doi:10.1109/ACEPT.2016.7811530.

Liu, C., Xu, X., and Hu, D. (2015). Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3), 385–398. doi:10.1109/TSMC.2014.2358639.

Lowe, R., Mordatch, I., Abbeel, P., Wu, Y., Tamar, A., and Harb, J. (2017). Learning to cooperate, compete, and communicate. URL https://openai.com/blog/learning-to-cooperate-compete-and-communicate/.

Miller, R. and Malinowski, J. (1994). *Power System Operation*. McGraw-Hill Education.

Moffaert, K.V. and Nowé, A. (2014). Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies. Technical report. URL http://www.jmlr.org/papers/volume15/vanmoffaert14a/vanmoffaert14a.pdf.

National Grid Electricity Transmission (2017). *The Grid Code*. URL https://www.nationalgrid.com/sites/default/files/documents/8589935310-Complete%20Grid%20Code.pdf.

Rozada, S. (2018). Frequency control in unbalanced distribution systems. *City, University of London MSc Data Science Thesis*.

Steitz, C. (2019). Nissan leaf gets approval for vehicle-to-grid use in germany. URL https://reut.rs/2OISFat.