A Decentralized Multi-Agent Energy Management Strategy Based on a Look-Ahead Reinforcement Learning Approach

Abstract

An energy management strategy (EMS) has an essential role in ameliorating the efficiency and lifetime of the powertrain components in a hybrid fuel cell vehicle (HFCV). The EMS of intelligent HFCVs is equipped with advanced data-driven techniques to efficiently distribute the power flow among the power sources, which have heterogeneous energetic characteristics. Decentralized EMSs provide higher modularity (plug and play) and reliability compared to the centralized data-driven strategies. Modularity is the specification that promotes the discovery of new components in a powertrain system without the need for reconfiguration. Hence, this paper puts forward a decentralized reinforcement learning (Dec-RL) framework for designing an EMS in a heavy-duty HFCV. The studied powertrain is composed of two parallel fuel cell systems (FCSs) and a battery pack. The contribution of the suggested multi-agent approach lies in the development of a fully decentralized learning strategy composed of several connected local modules. The performance of the proposed approach is investigated through several simulations and experimental tests. The results indicate the advantage of the established Dec-RL control scheme in convergence speed and optimization criteria.

Introduction

Heavy-duty hybrid fuel cell vehicles (HFCVs) are recognized as promising candidates to alleviate concerns regarding the growing greenhouse gas emissions [1-3]. Proton exchange membrane fuel cell (PEMFC) has a good reputation as the most appropriate option for this application due to its power density, low-temperature operation range, low noise, and high-efficiency performance [4]. Since the heavy-duty HFCVs need a high-power level, their powertrain systems are usually composed of a multi-stack fuel cell system (MFCS), as shown in Figure 1. In [5], a detailed survey of the MFCSs with different fluidic and power conditioning architectures is provided. In the heavy-duty HFCVs, the energy management strategy (EMS) unit is a crucial control part to reduce the total end-user costs and meet the powertrain components' requirements [2]. Several studies have been done in the literature on designing the EMSs for the HFCVs with an MFCS structure [6-13].

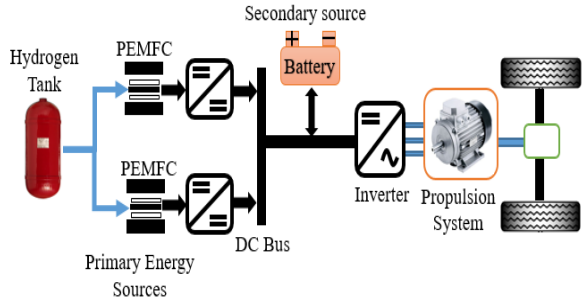

Figure 1. Heavy-duty HFCV with an MFCS and a battery pack.

However, one missing point in the previous EMSs is that they did not consider the time-varying characteristics of the powertrain components and operation in various driving profiles, leading to inaccurate optimal solutions for the developed EMSs. For instance, the maximal power of an FCS decreases by 10% during the lifetime, and this value is not reached with the same current at the beginning and the end of life [14]. To tackle this challenge, the model-free type of reinforcement learning (RL) algorithm [15] as a powerful data-driven approach has recently attracted much attention in different real-world applications, such as diesel engines [16] and electric vehicles. RL-based approaches are starting to show some accomplishments in real-world applications. At the same time, these approaches are faced with a set of challenges [17]. Among these challenges, operation safety is an important one which has been the focus of several research works. Some of the proposed approaches in this regard are coach-actor-double critic [18], learning-based model predictive control [19], robust model predictive control [20], parallel constrained policy optimization [21], shielding [22, 23], Lyapunov-based [24]. In [25], the RL-based EMSs of hybrid electric vehicles (HEVs) and plug-in hybrid electric vehicles (PHEVs) are reviewed, and various potential DRL-based solutions are presented. An RL-based EMS is designed to split the requested power between a battery pack and an engine-generator set in [26] with the aim of minimizing the fuel consumption over different driving profiles. Moreover, a transition probability matrix is introduced to learn a specific driving schedule. A fuzzy encoding and nearest neighbour methods are suggested in [27] to achieve a predictive RL-based EMS for an HEV. In order to learn the transition probabilities and statistical information, a finite-state Markov chain is utilized. In [28], a Kullback–Leibler (KL) divergence rate is proposed to decide when the transition probability matrix adaptation is vital in the real-time application. An investigation between one-step (Q-learning) and multi-step (Dyna and Dyna-H) algorithms is conducted in [29]. In [30], a fast Q-learning is proposed to achieve an online updating EMS and speed up the control policy's convergence rate. Furthermore, cloud-based computation is employed to address the computation burden [31]. This paper studies a model-free predictive EMS for a connected electrified off-highway vehicle. Several multi-step RL-based EMS is proposed to enable the all-life-long online optimization. In [32], a model-free supervisory EMS via a double Q-learning approach is developed for the HEVs with the charge-sustaining scenarios. An ensemble RL algorithm based on Q-learning method is

proposed to form a centralized multi-agent system with different state combinations [33]. An analysis is presented in [34] regarding the adaptability interpretation with different real-world driving conditions, such as driving cycle, load, road grade, and traffic. In [35], a DRL-based EMS is proposed to learn from the environment and make an appropriate decision without any prediction. An actor-critic RL by a deep deterministic policy gradient (DDPG) method is offered in [36] to incorporate the traffic information and passengers' number into the EMS design. A double deep Q-learning algorithm is proposed in [37] to improve the convergence rate and optimization performance. The suggested approach prevents the training process from falling into the over-optimistic estimate of policy value. To address the training time concern of the DRL algorithm, a bi-level adaptive EMS is constructed by integrating a transfer learning (TL) method [38]. Even though many advanced EMSs have been developed in the literature, few studies have considered the constrained setting for the training safety issue. To ensure safety, an coach-actor-double critic approach is introduced in [18]. In this work, when the actor's output exceeds the feasible solutions, the coach will be in charge of the EMS. A knowledge transfer among four types of HEVs is studied in [39], and the convergence efficiency has been improved. In [40], a rule-based controller is combined with a DRL algorithm to eliminate irrational torque allocation. Furthermore, a hybrid experience replay method is utilized to address the disturbance sensitivity. In [41], the power distribution history data is considered as an expert driver knowledge to guide a DRL algorithm to design a human-like EMS.

For the first time in the HFCV applications, to achieve an adaptive optimal EMS, a radial-basis neural network using the RL framework is introduced by Lin *et al.* [42]. The proposed EMS did not need any prior knowledge of future driving cycles. In another study, Hsu *et al.* [43] established an RL EMS for an FCS/Battery hybrid electric vehicle. Yuan *et al.* [44] proposed a hierarchical RL control strategy with real-time capability. Reddy *et al.* [45] suggested an RL EMS that can autonomously learn the optimal policy using a powertrain system. To minimize the final cost of a plug-in HFCV, Lin *et al.* [46] put forward an online recursive RL approach. Sun *et al.* [47] developed a hierarchical multi-objective RL strategy by merging transition probability matrix into equivalent consumption minimization (ECMS). The majority of the learning approaches are founded on centralized EMSs (Cen-EMS). Due to their inherent concentrated characteristic, these control strategies are vulnerable to electrical fault or malfunction of the power sources, especially for the high-power heavy-duty HFCVs with coupled MFCSs. Additionally, these Cen-EMSs may lose their operational performance in case of adding or removing one FCS. In intelligent EMSs, large volumes of data collected from intelligent FC modules can be used as preliminary information for the EMSs. Additionally, efficient utilization of massive information gained from vehicle-to-vehicle (V2V) and vehicle-to-everything (V2X) is a promising approach for improving EMSs performance. However, the training procedure of the data-driven EMSs with these massive volumes of information may lead to significant computational complexity and even yield diverging optimization results. On top of that, the FC modules' aggregated data at the central control unit may be easily exposed to potential cyberattacks. As a result, there is a need to advance these Cen-EMSs in durability, modularity (plug and play), and computational complexity. One promising solution is to develop a modular FCS with a decentralized EMS (Dec-EMS) configuration [48]. As shown in (Figure 2 (b)), compared to the Cen-EMS (Figure 2 (a)) with a central control unit, the Dec-EMS consists of several connected control modules that simultaneously solve the EMS optimization problem.

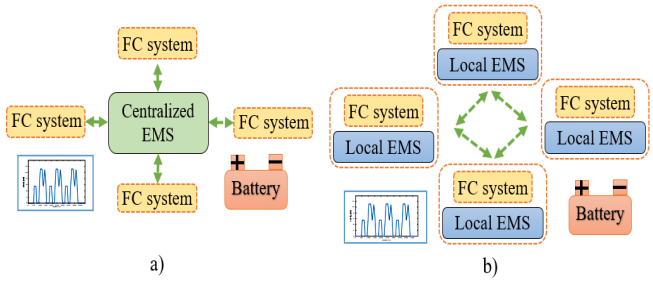

Figure 2. The powertrain EMS architectures: a) the centralized controller unit (Cen-EMS), b) the fully decentralized controller (Dec-EMS).

The multi-agent RL approaches have several advantages over the single-agent RL methods [49-51]. First and foremost, a fast learning performance can be achieved in the multi-agent RL configuration thanks to the parallel computation. In this way, the agents exploit the power allocation objective in a decentralized form. The multi-agent RL algorithms with similar tasks can reach superior performance and learn faster utilizing the experience-sharing concept. Since information analysis is conducted locally, the amount of exchanged data is also limited. This feature reduces the need for an expensive network and decreases issues like synchronization and delay. Moreover, when one agent stops the regular operation due to electrical malfunction or cyberattacks, the rest can compensate for the absence of the failed agent. In this regard, the multi-agent RL approaches are inherently robust. This configuration also offers easy insertion of new agents, leading to a high degree of scalability and flexibility. This plug-and-play characteristic allows manufacturers to embed a reconfigurable controller, which reduces the final cost of installing new powertrain components.

In this study, decentralized reinforcement learning (Dec-RL) is proposed to tackle the EMS optimization problem of an HFCV with an MFCS configuration. In contrast to the centralized RL (Cen-RL), where the collected data by all of the power components are analyzed at a single control unit, the decentralized data-driven framework consists of several multi-agent units that solve the EMS optimization problem through cooperation.

To the best of the authors' knowledge, this paper is the first attempt to put forward a Dec-RL framework for the EMS optimization problem of a heavy-duty HFCV powered by multi-FC modules as the primary sources and a battery pack as the secondary one.

The remainder of this paper is organized in the following way. In section II, the powertrain components and modeling are presented in detail. A general overview of the RL framework is carried out in Section III. Section IV presents the mathematical formulation of the proposed Dec-RL EMS. In section V, the simulation results' analyses are illustrated to investigate the performance of the proposed strategy. Section VI provides the implementation results. Lastly, Section VII summarizes the significant findings.

Powertrain components and modeling

General powertrain structure

To analyze the performance of the suggested Dec-RL EMS, a modular test bench based on a heavy-duty HFCV [52, 53] is established, as shown in Figure 3 and Figure 4. This small-scale test bench consists of two FC modules, a battery pack, a programmable electronic load, and multi-range programmable DC power supplies for simulating the load profiles. Each FC module's key component is a 500-W open-cathode PEMFC (*Horizon, model: H-500*), a smoothing inductor, and an adjustable unidirectional boost DC-DC converter (Zahn Electronics™, model: DC5036-SU). Six series 12-V 18-Ah battery packs supply the voltage of the DC bus. Each FC module has its autonomous Dec-EMS inside of a National Instrument CompactRIO (NI 9022). CompactRIO is a control unit for industrial applications with the capability to run under harsh and noisy conditions. It is composed of an FPGA, a real-time controller, reconfigurable I/O, and an Ethernet chassis. FPGA is utilized for measuring the data and to execute highly efficient data processing. The real-time controller contains a microprocessor unit for implementing the control strategies. During the deployment process, the LabVIEW software program has been compiled for the real-time controller and FPGA. The optimal reference of each FC module is calculated at every control instant with an interval of 10 Hz.

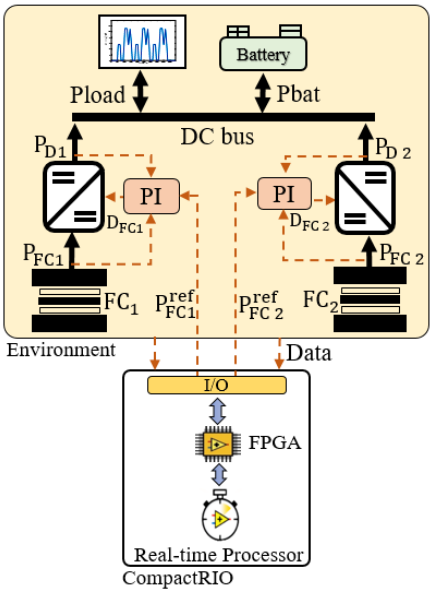

Figure 3. The test-bench configuration involving two FC modules and one battery unit.

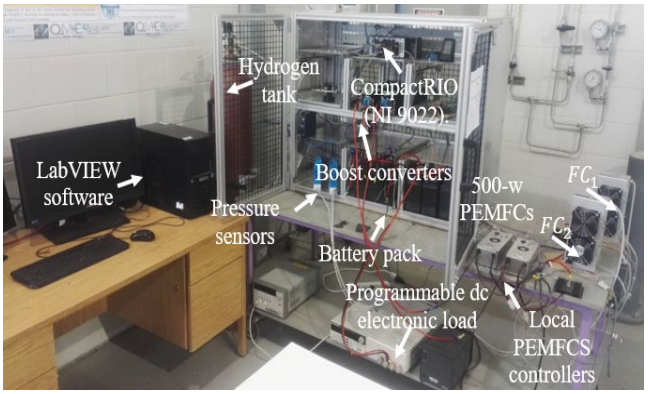

Figure 4. The developed small-scale modular FCS test bench.

The power equilibrium of the FC modules and the battery unit is formulated in (1).

$$\sum_{m=1}^{M} P_m\,[k] D_m[k] + P_B[k] = P_L[k], \qquad (1)$$

where $P_{\mathrm{m}}, m \in M, M = \{1,2\}$ denotes the generated power by each of the 500-W FCSs, D_{m} is the control signal of the boost converters, P_B is the power provided by the battery unit, P_L is the requested power from the propulsion system, and k stands for each simulation moment.

FCS modeling and constraints

In this work, each of the 500-W FCSs, FC_m, is modeled as a voltage source where their polarization curves and hydrogen mass flows versus requested currents are described by experimentally validated quasi-static curves, as shown in Figure 5. The technical data of the utilized FCSs are reported in Table 1.

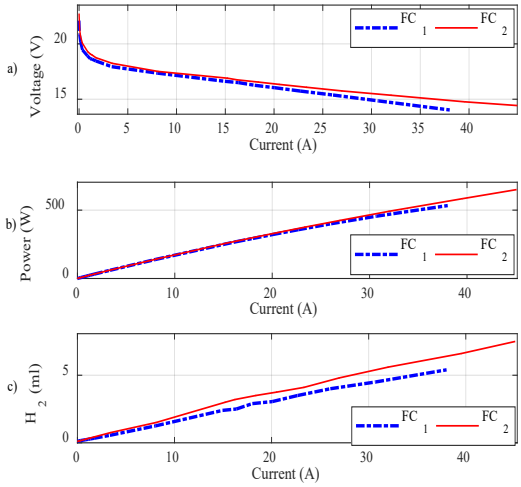

Figure 5. The characteristic curves of the two real FCS modules: (a) polarization curves, (b) power curves, and (c) hydrogen curves.

Table 1. Technical specification of the open-cathode PEMFC stacks (H-500).

PEMFC Stack Information	
Number of cells= 24	Hydrogen pressure= 0.45-0.55bar
Rated power= 500W	External temperature= 5-30°C
Rated performance= 14.4V @35A	Max stack temperature= 65°C
Output voltage range= 12V-24V	Humidification= Self-humidified
Reactants= Hydrogen and Air	Cooling= Air (cooling fan)
Flow rate at max output= 6.5L/min	Efficiency of stack= 40% @ 14.4V

Equation (2) enforces power and slew rate limits.

$$P_{m,min} \leq P_m[k] \leq P_{m,max}, \quad (2.a)$$

$$R_{d,m}\Delta t \leq P_m[k] - P_m[k-1] \leq R_{u,m}\Delta t, \quad (2.b)$$

where $P_{m,min}$ and $P_{m,max}$ are the minimum and maximum values for P_m, $R_{d,m}$ and $R_{u,m}$ are boundaries of the slew rate, and Δt indicates the time step. As explained in [54], for rising, a dynamic limitation of 10% of the maximum power per second, and for falling, 30% of the maximum power per second have been considered for the PEMFC stack operation. These constraints prevent the FCS from sudden changes, which can result in degradation.

Battery modeling and constraints

The first-order RC model of the battery pack is formulated by

$$I_B[k] = \frac{V_0[k] - R_s I_B[k] - V_B[k]}{R_c} + C_c \frac{d}{dt}(V_0[k] - R_s I_B[k] - V_B[k]), \quad (3)$$

where I_B is the battery pack current, V_0 is the open-circuit voltage, R_s is the series ohmic resistance, V_B is the output terminal voltage, R_c denotes the polarization resistance, and C_c is the polarization capacitor. Equation (4) imposes power and slew rate limits.

$$P_{B,min} \leq P_B[k] \leq P_{B,max}, \quad (4.a)$$

$$R_{d,B}\Delta t \leq P_B[k] - P_B[k-1] \leq R_{u,B}\Delta t, \quad (4.b)$$

where $P_{B,min}$ and $P_{B,max}$ are the minimum and maximum limits of P_B, respectively, and $R_{d,B}$ and $R_{u,B}$ are the slew rate boundaries of P_B. Equation (5) presents the state of charge (SoC) calculation formula along with the constraints on the battery SOC level.

$$SoC[k+1] = SoC[k] - \frac{P_B[k]\Delta t}{Q_B V_B[k] 3600}, \quad (5.a)$$

$$SoC_{min} \leq SoC[k] \leq SoC_{max}, \quad (5.b)$$

$$SoC[0] = SoC_0, \quad (5.c)$$

where SoC_{min} and SoC_{max} denote the minimum and maximum limits of SoC, respectively, SoC_0 is the initial SoC level, and Q_B represents the battery capacity. The battery lifetime is affected by the depth of discharge (DOD) and is defined *as* an initial capacity drop (reaching 80% of the initial capacity). The state of health (SoH) is calculated by

$$SoH[k+1] = SoH[k] - \frac{|P_B[k]|\Delta t}{2 n_B Q_B V_B[k] 3600}, \quad (6.a)$$

$$SoH_{min} \leq SoH[k], \quad (6.b)$$

$$SoH[0] = SoH_0, \quad (6.c)$$

where SoH_{min} and SoH_0 indicate the minimum and initial SoH, respectively, and n_B is the total number of cycles during the whole lifetime of the battery pack. The obtained parameters of the battery pack have been obtained from experimental tests and reported in Table 2.

Boost converter modeling and characteristics

The two converters are modeled as follows:

$$L_m \frac{d}{dt} I_m[k] = V_m[k] - V_{h,m}[k] - r_m I_m[k],$$

$$\begin{cases} V_{h,m}[k] = m_{h,m} V_B[k] \\ I_{h,m}[k] = m_{h,m} I_m[k] \eta_{h,m}^{z} \end{cases} \quad z = \begin{cases} 1, & \text{if } P_m > 0 \\ -1, & \text{if } P_m < 0 \end{cases} \quad (7)$$

where I_m and V_m are the current and voltage of FC_m, respectively, L_m presents the smoothing inductor inductance, r_m is the smoothing inductor resistance, $\eta_{h,m}$ is the average efficiency, and $m_{h,m}$ is the modulation ratio of the converters. The estimated parameters of the boost converters are presented in Table 2.

Table 2. The approximated battery and converter parameters.

$V_0 = 12.21\,V$	$Q_B = 18.2Ah$	$L_m = 1.1$ mH	$r_m = 23.9$ mΩ
$R_s = 0.0141\Omega$	$R_c = 0.0177\Omega$	$\eta_{h,m} = 96.21\%$	

General description of the reinforcement learning structure in power allocation strategy application

A Markov decision process (MDP) is defined as a tuple ($\mathcal{S}$, $\mathcal{A}$, P, R, T), where $\mathcal{S}$ and $\mathcal{A}$ denote the set of states $s \in \mathcal{S}$ (the requested power, SOC) and the control actions $a \in \mathcal{A}$ (the FC modules' output power), respectively. $P(s_{t+1} | s_t, a_t): \mathcal{S} \times \mathcal{A} \rightarrow P(s)$ is the probability of transitioning into $s_{t+1} \in \mathcal{S}$ at time $t+1$ when an agent (the EMS control unit) takes an action $a_t \in \mathcal{A}$, in the state $s_t \in \mathcal{S}$ at time t. $R_t(s_t, a_t, s_{t+1}): \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the reward value obtained when an action $a_t \in \mathcal{A}$ is taken. T presents a finite time horizon that the MDP is solved to seek a control policy $\pi_\theta \in \Pi$. The control policy relies on the value of $Q(s_t, a_t)$ and defines an action $a \in \mathcal{A}$ that must be taken in each state $s_t \in \mathcal{S}$ for maximizing the discounted cumulative rewards $E[\sum_{i=0}^{t} \gamma^i R_{t+1+i} | s = s_t, a = a_t]$, where $\gamma \in [0,1]$ is a discounting factor. The primary purpose of the Q-learning algorithm is to determine the optimal control strategy $\pi_\theta^* = argmax\ Q(s_t, a_t)$ that maximizes the Q-value.

The mathematical formulation of the proposed decentralized reinforcement learning EMS framework

In this section, the process of developing the Cen-EMS and the Dec-EMS approaches for the MFCV EMS optimization problem is developed. The central concept of a multi-agent RL framework for the EMS optimization problem of a heavy-duty MFCV is shown in Figure 6. Generally, in the multi-agent RL approach, several agents communicate with the environment to determine how to develop the control policies and take optimal actions in the presence of unwanted disturbance. The environment is formed from the powertrain components, driving profile, satellites, global position system (GPS), V2V&V2X, and traffic information.

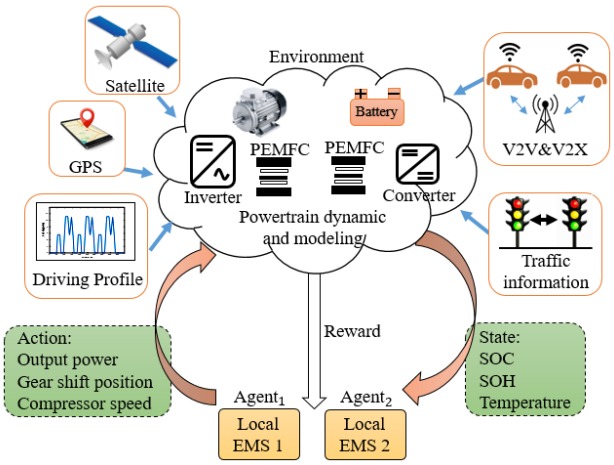

Figure 6. The general decentralized multi-agent reinforcement learning.

The proposed multi-agent EMS is composed of a set of FC modules and a battery pack. Each FC module agent aims to decrease the total cost while taking into account the powertrain operation characteristics. Q-learning [55] is nominated for each FC module agent's learning process, and all FC modules are considered to run their Q learning algorithms synchronously. The corresponding RL agents optimally allocate the FC modules' output powers in a discrete action space in the following manner. Firstly, each control unit uploads the optimal action to the powertrain system and receives a reward value. Secondly, each FC module EMS updates its local control policy through the received reward value. A sufficient number of the training process is iteratively executed until all FC modules' control policies converge. After completing the learning process, each local EMS has its fine-tuned local control policy, which is more general to new unobserved states. In the following subsections, the mathematical equations related to the Cen-RL and the Dec-RL approaches are thoroughly presented.

The centralized reinforcement learning approach

The mathematical formulation of the centralized EMS is formulated as follows.

1) State-space: The state space of the Dec-RL agent for the FC modules is determined by

$$\mathcal{S}_{FC} = \{P_L[k], SoC[k]\} \quad (8)$$

where k represents the planning time of the two FC modules and the battery pack, $P_L[k]$ denotes the P_L value at time k, $SoC[k]$ is the SOC of the battery pack at time k, which is calculated using equation (5.a) depending on the generated powers of the FC modules at time k ($P_m[k]$) and $P_L[k]$.

2) Action space: The optimal output power (action) for each FC module impacts on the powertrain (environment), including the current state, $\mathcal{S}^{FC}$ as presented in (8). The action space of the FC modules is demonstrated below.

$$\mathcal{A}_{FC} = \{P_m[k])\} \quad (9)$$

where $P_m[k]$ denotes the generated powers of the FC modules at time k. $P_m[k]$ must take into account (2.a) and (2.b).

3) Reward function: The multi-objective reward function is developed as the sum of the negative hydrogen and degradation expenses associated with the FC modules and the battery unit and their operating characteristics. The reward function for the FC modules' agent, $\mathcal{R}_T[k]$, is defined as

$$\mathcal{R}_T[k] = -1 \times (\sum_m S_{H,m}[k] + S_{d,m}[k] + S_B[k] + S_{SoC}), \quad (10.a)$$

The hydrogen cost of each FC module, $S_{H,m}$, is computed by

$$S_{H,m}[k] = h_m[k] C_{H_2} \Delta t, \quad (10.b)$$

where h_m is the hydrogen consumption, C_{H_2} is the hydrogen price, and Δt indicates the time step. Each FC module degradation cost, $S_{d,m}$, is calculated by

$$S_{d,m}[k] = S^l_{d,m}[k] + S^h_{d,m}[k], \quad (10.c)$$

where $S^l_{d,m}$ and $S^h_{d,m}$ are the related expenses to low-power and high-power operation, respectively. $S^h_{d,m}$ is calculated by

$$S^l_{d,m}[k] = \frac{n_m \varepsilon_l C_{FC,m} \Delta t \mu_{l,m}}{3600\, V_{n,m}}, \quad (10.d)$$

$$S^h_{d,m}[k] = \frac{n_m \varepsilon_h C_{FC,m} \Delta t \mu_{h,m}}{3600\, V_{n,m}}, \quad (10.e)$$

where ε_l and ε_h are the low-power and high-power cell degradation rates, respectively. The values of these variables are given in Table 3, which are adopted from [56, 57].

Table 3. Cell degradation rate.

Variable	Symbol	Value	Unit
Low-power	ε_l	8.662	$\mu V/h$
High-power	ε_h	10	$\mu V/h$

where n_m represents the number of cells in each FC_m, $V_{n,m}$ is 10 % of the nominal FC_m voltage drop, and C_{FC_m} is FCS cost. $\mu_{l,m}$ and $\mu_{h,m}$ are equal to

$$\mu_{l,m} = \begin{cases} 1, & \text{if } P_{min,m} \le P_m[\mathrm{k}] \le 0.2 \times P_{nom,m} \\ 0, & \text{otherwise.} \end{cases}, \quad (11.a)$$

$$\mu_{h,m} = \begin{cases} 1, & \text{if } 0.8 \times P_{nom,m} \le P_m[k] \le P_{max,m} \\ 0, & \text{otherwise.} \end{cases}, \quad (11.b)$$

where $P_{nom,m}$ is the output power recommended by the FCS manufacturing company for nominal use of FCS [57]. The battery degradation cost, S_B, is determined by

$$S_B[k] = C_B(SoH_B[k] - SoH_B[0]), \quad (11.c)$$

where C_B is the battery price. The punishment term, S_{SoC}, is used to measure the SoC level variation, which is defined by

$$s_{SoC}[k] = \beta(SoC[k] - SoC_0)^2, \quad (11.d)$$

where SoC_0 is the initial SoC, and β is a significant positive coefficient. Table 4 demonstrates the reference price of hydrogen, battery, and the FCS.

Table 4. The reference price of hydrogen, battery, and FCS.

Cost	Symbol	Value	Unit
Hydrogen	C_{H_2}	3.9254 [58]	\$/Kg
FCS	C_{FC_m}	35 [59]	\$/kW
Battery unit	C_B	189 [60]	\$/kWh

The main structure with the components of the developed Cen-RL EMS is presented in Figure7.

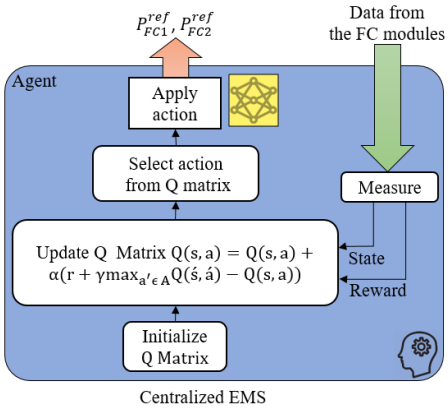

Figure 7. The Cen-RL algorithm configuration.

The decentralized reinforcement learning approach

The mathematical formulation of the decentralized learning-based algorithm is presented below.

1) State-space: The state space of the Dec-RL agents for the FC modules is determined by

$$\mathcal{S}_{FC}^{m} = \{P_L[k], SoC[k]\}. \quad (12)$$

2) Action space: The action space of the FC modules is demonstrated below.

$$\mathcal{A}_{FC}^{m} = \{P_m[k])\} \quad (13)$$

3) Reward function: The reward function for the FC modules' agents, $\mathcal{R}_T^m$, is formulated as

$$\mathcal{R}_T^m = -1 \times (S_{H,m}[k] + S_{d,m}[k] + S_B[k] + S_{SoC}), \quad (14.a)$$

The hydrogen cost of each FC module, $S_{H,m}$, is computed by

$$S_{H,m}[k] = h_m[k] C_{H_2} \Delta t. \quad (14.b)$$

Each FC module degradation cost, $S_{d,m}$, is calculated by

$$S_{d,m}[k] = S_{d,m}^{l}[k] + S_{d,m}^{h}[k], \quad (14.c)$$

where $S_{d,m}^{l}$and $S_{d,m}^{h}$are expenses of low-power, high-power, respectively. The battery degradation cost, S_B, is determined by

$$S_B[k] = C_B(SoH_B[k] - SoH_B[0]). \quad (15.a)$$

s_{SoC} is a punishment item to measure the SoC level variation, which is defined by

$$s_{SoC}[k] = \beta(SoC[k] - SoC_0)^2. \quad (15.b)$$

The decentralized structure of the Cen-RL EMS and the pseudo-code are presented in Figure 8 and Algortim.1.

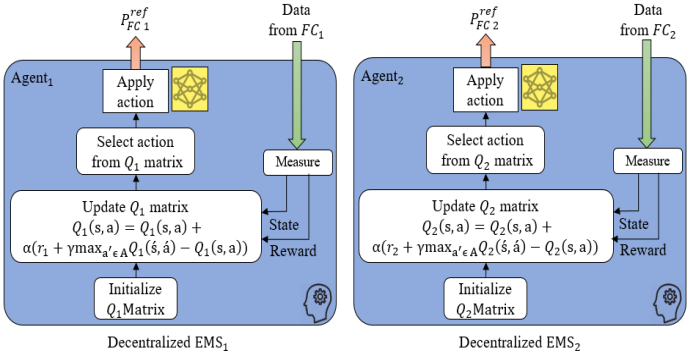

Figure 8. The configuration of the Dec-RL EMS involving N-connected agents.

Algorithm.1: Dec-RL process for the power allocation strategy problem.

1. Initialize each fuel cell module's constraints, the battery, output power, and operation characteristics.
2. Initialize the local Q_n matrix, iteration number, learning parameters, the state, and action lists of each agent
%%Seeking an optimal power allocation policy for each FC modules' agent
3. **For** local learning episode = 1, Maximum value do
4. **For** the selected load profile time step=1, Maximum value do
5. Select an action $a_t \in \mathcal{A}$ with $P(s_{t+1} | s_t, a_t)$ for each state $s_t \in \mathcal{S}$ (ε-greedy policy) and observe $R_t(s_t, a_t, s_{t+1})$
6. update $Q_n(s_t, a_t)$ by
$Q_n(s,a) = Q_n(s,a) + \alpha(r_n + \gamma \max_{a' \epsilon A} Q_n(\acute{s}, \acute{a}) - Q_n(s,a))$
7. **end**
8. **end**

Simulation results analysis and discussion

In this section, the performance of the proposed Dec-RL EMS scheme is investigated. Firstly, the suggested multi-agent method is compared with the Cen-RL and dynamic programming (DP). Secondly, the learning-based strategies are improved from a single-step optimization algorithm to a multi-step moving horizon one. The computational complexity is thoroughly influenced by the PC hardware (Processor unit= Core i5, 2.30 GHz, RAM= 4.00 GB). During the evaluation phase, the Q matrix is initialized with zero values (cold-start). In the learning-based optimization approaches, the learning rate (α), the random action exploration rate (ε), and discount factor (γ), are 0.3, 0.999, and 0.1, respectively. The learning parameters are customized according to the primary learning sequences based on the system specifications [61, 62]. In this regard, several values are tried, and the ones with the best performance are selected. For instance, using a large value for the learning rate (α) would result in a lot of oscillations and local optima while a small learning rate could lead to a learning process. Similar to the learning rate, the discount factor (γ) can be adjusted based on initial learning results. The discretization values of the state space and the control actions are listed in Table 5.

Table 5. The state-space variables and the action space variables discretization.

Variable	Lower band	Upper band	Discretization
P_L	Minimum value	Maximum value	15
SoC	0.6	0.8	20
P_1	0.125×Maximum power	Maximum power	10
P_2	0.125×Maximum power	Maximum power	10

To evaluate the designed EMS performance, a real driving profile [52] from a heavy-duty HFCV is used throughout the training procedure, as demonstrated in Figure 9. Since the maximum power of the driving profile is higher than the limitations of the modular small-scale test bench, the driving profile has been scaled down.

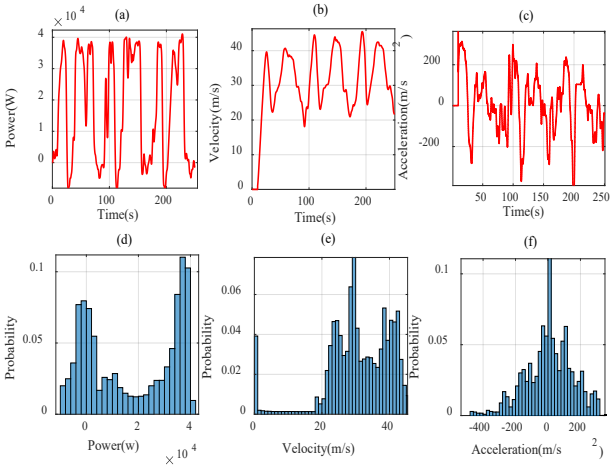

Figure 9. The real driving profile characteristics: (a) power, (b) velocity, (c) acceleration, (d) power distribution, (e) velocity distribution, and (f) acceleration distribution.

General performance investigation

The optimized output power trajectories and the SoC evaluation obtained by the Cen-RL and the Dec-RL EMSs after 2000 times iteration are shown in Figure.10 and Figure 11. As it can be seen, the operating points of the FC modules under the Cen-RL and Dec-RL EMSs are primarily located in the high efficiency and low hydrogen-consumption areas. The SoC trajectories of the battery unit under Cen-RL and Dec-RL EMSs fluctuate between 69% and 71%, and both SOC oscillations are almost similar.

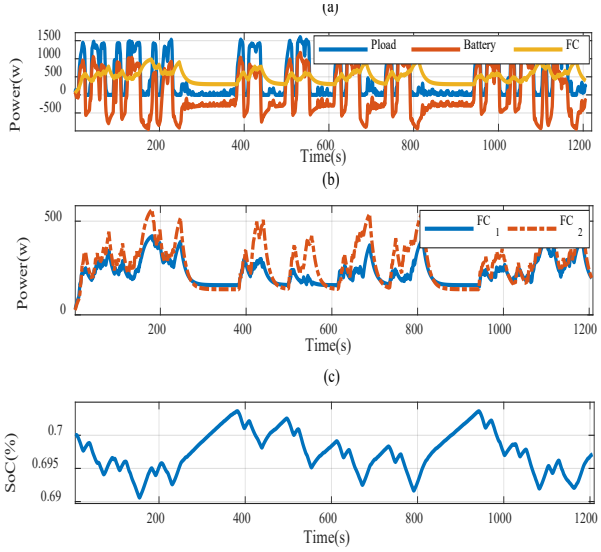

Figure 10. Optimal EMSs results based on Cen-RL (Pload: the requested power, FCs: the power provided by the FC modules (FC_1, FC_2), Battery: the battery power): (a) the power profiles, (b) the power profiles of the FC modules, and (c) the SoC level trajectory.

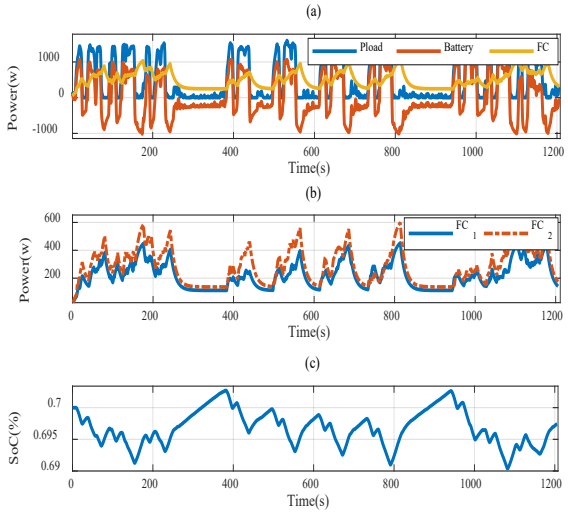

Figure 11. Optimal EMSs results based on Dec-RL (Pload: the requested power, FCs: the power provided by the FC modules (FC_1, FC_2), Battery: the battery power): (a) the power profiles, (b) the power profiles of the FC modules, and (c) the SoC level trajectory.

A comparison between the FC modules' power profiles in the time domain and power distribution under Cen-RL and Dec-RL approaches is shown in Figure 12 and Figure 13. It can be viewed that the FC modules collaborate at the efficient regions to meet the load power profile and reduce the total reward functions. Nevertheless, because of the FC modules' slow response characteristics, the battery unit supplies the peaks and fast dynamic powers. In case of gradual ageing phenomena or employing two FCSs with different degradation levels, it is crucial to update the hydrogen consumption, polarization curves, and degradation levels. One of the promising methodologies to adapt to these changes is to integrate the proposed framework with an online identification method to capture the actual fuel cell characteristics [63]. From the viewpoint of a PAS, different FC characteristics means different feasible search space for the optimization problem (e.g. $P_{FC,Min} \leq P_{FC} \leq P_{FC,Max}$). Therefore, the extracted power from each FC by the PAS is different since the feasible search space of each FC is different.

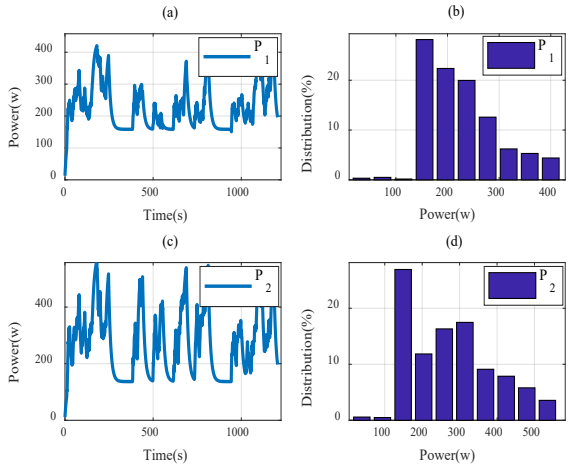

Figure 12. The optimized power profile and the power distribution of the FC modules utilizing Cen-RL (a) the power profile of the FC_1, (b) the

distribution of the FC_1, (c) the power profile of the FC_2, and (d) the distribution of the FC_2.

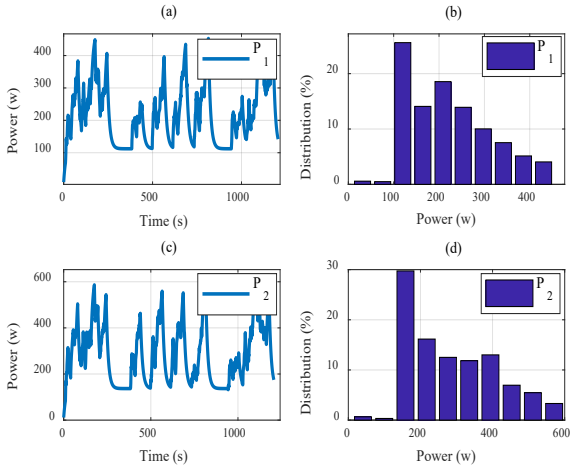

Figure 13. The optimized power profile and the power distribution of the FC modules employing Dec-RL (a) the power profile of the FC_1, (b) the distribution of the FC_1, (c) the power profile of the FC_2, and (d) the distribution of the FC_2.

A comparison between the learning curves of the Cen-RL and Dec-RL total rewards is presented in Figure 14. For both RL-based approaches, the training processes are repeated ten times. It should be mentioned that the training curves increase and converge as the optimal control policies for the FC modules and the battery unit learn from the reward functions. 783 learning iterations are required for Cen-RL to reach the optimal policy, while Dec-RL only needs 622 learning iterations, a 20.35% iterations reduction. It can be inferred that the Dec-RL EMS accelerates the learning process of the optimal policy as a consequence of the reduction of the Q matrix and the cooperation of the FC modules and the battery pack.

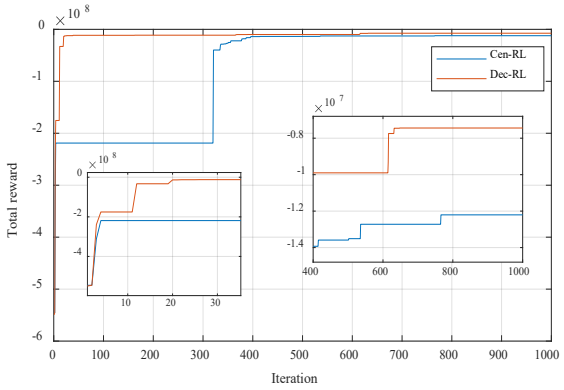

Figure 14. The evolution of the total rewards reached during the learning step of the Cen-RL and Dec-RL approaches.

The final price (after SOC correction) and the computational time are listed in Table 6. For the DP approach, the computation time is the total required time of performing the optimization for the entire driving cycle. Concerning the RL-based strategy, the computation time refers to the required time of the training process. The proposed learning-based approaches have shown a very close total cost to the DP algorithm. The total costs based on Cen-RL and Dec-RL are \$0.051and \$0.0471($FC_1 + FC_2$), respectively, which is 13.33% and 4.44% higher than DP (\$0.045). In terms of computational time complexity, the proposed Dec-RL method has been much faster than others because of the decentralized structure. The computational burden has been reduced by 66.66% and 92.61% concerning Cen-RL and DP, respectively.

Table 6. The comparison analysis of the developed algorithms.

	DP	Cen-RL	Dec-RL	
	$FC_1 + FC_2$	$FC_1 + FC_2$	FC_1	FC_2
Computational time (s)	205	45.460	15.1542	15.286
Number of iteration	-	783	621	621
Total cost (\$)	0.045	0.051	0.0235	0.0236

Analysis of moving finite learning ahead horizon

This subsection scrutinizes how multi-step ahead RL-based optimization influences the optimal cost and the convergence speed of the Cen-RL and Decen-RL algorithms. Since velocity prediction is not in the scope of this study, the velocity profile is supposed to be precisely known. The Cen-RL and Decen-RL EMSs with two steps to 26 steps are compared in terms of the convergence speed and the total reward. The associated results with the centralized and decentralized algorithms are demonstrated in Figure 15 and Figure 16.

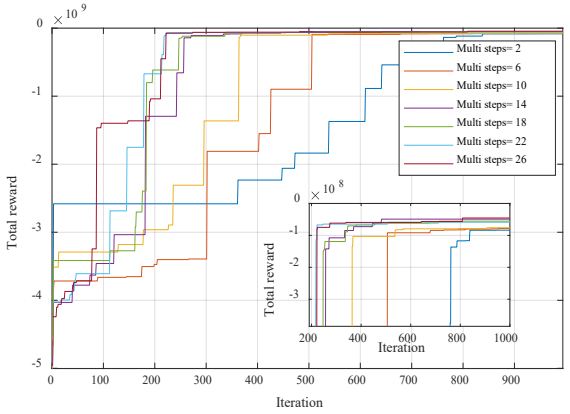

Figure 15. The convergence speeds and the total negative rewards comparison of the Cen-RL trajectories with different multi-step prediction values.

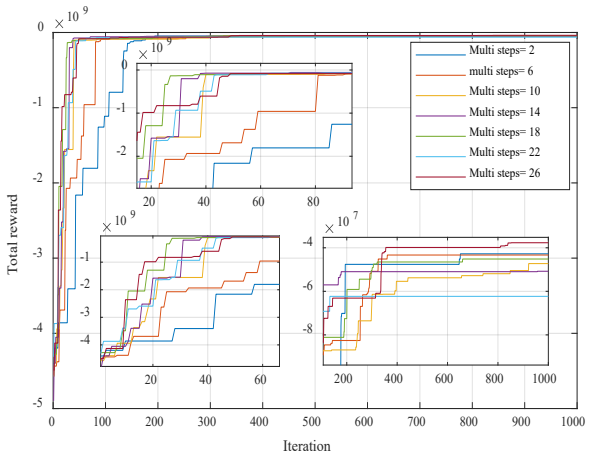

Figure 16. The convergence speeds and the total negative rewards comparison of the Dec-RL trajectories with different multi-step prediction values.

It becomes apparent that convergence speed increases, and the total final cost reduce with adding moving finite learning ahead horizons into both learning-based methods. It can be concluded that it is vital to include a moving multi-step prediction for the real-time perspective. The optimal power trajectories with 26 steps ahead prediction are illustrated for the Cen-RL and Dec-RL EMSs in Figure 17 and Figure 18. As it can be seen, in both Cen-RL and Dec-RL with the moving horizon methods, the final SoC values converge exactly to 0.7, while in the previous section results, the data-driven methods could not finish in the same SoC values.

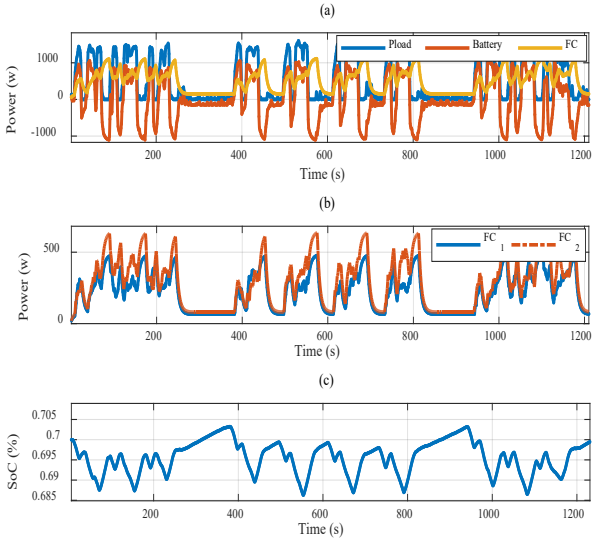

Figure 17. Optimal EMSs results based on Cen-RL with 26 steps (Pload: the requested power, FCs: the power provided by the FC modules (FC_1, FC_2), Battery: the battery power): (a) the power profiles, (b) the power profiles of the FC modules, and (c) the SoC level trajectory.

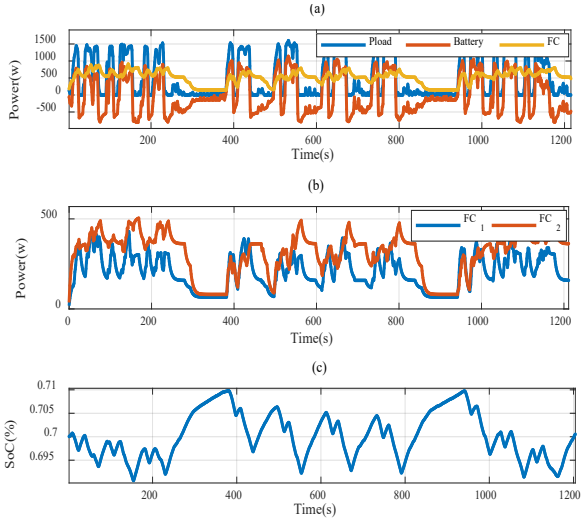

Figure 18. Optimal EMSs results based on Dec-RL with 26 steps (Pload: the requested power, FCs: the power provided by the FC modules (FC_1, FC_2), Battery: the battery power): (a) the power profiles, (b) the power profiles of the FC modules, and (c) the SoC level trajectory.

Experimental validation in the developed small-scaled test bench

In real-time applications, the computation time achievability of the decentralized data-driven approach is pivotal. The proposed Dec-RL EMS is implemented on the small-scale test bench to evaluate the previous simulation outcomes. The power profiles and SOC trajectories are illustrated in Figure 19. The FC modules operate in high efficiency and low-degradation zones. The total cost of Dec-RL ($0.049) is 10.90% lower than that of the Cen-RL algorithm ($0.055). The computational complexity of Dec-RL is about 49.78% lower than Cen-RL. The Dec-RL shifts the computational burden to the local controller units. In this way, a multi-agent platform with low-cost and limited computational capability units is sufficient. Furthermore, the multi-agent system leads to a reduced final cost. It can be inferred that in comparison to the centralized-based RL EMS, the suggested decentralized scheme is more cost-saving while having the ability to be implemented in real-time. Generally, the degradation terms have to be determined employing several long-duration ageing experimental tests. In this work, since the only usage of the degradation terms is to test the proposed PAS under performance attenuation conditions, both degradation rates are adopted from previously published papers.

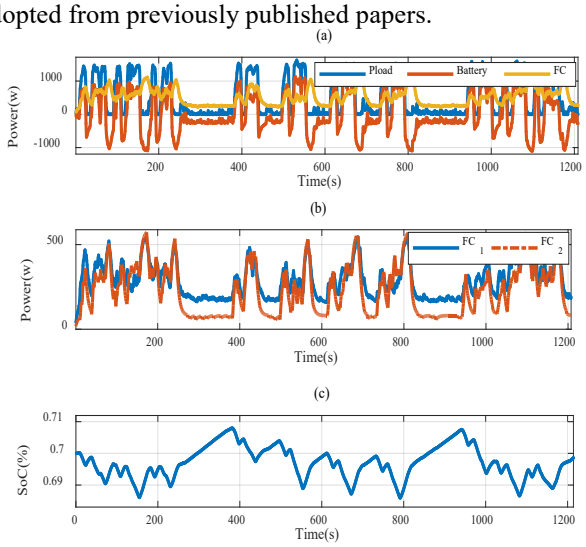

Figure 19. The implementation outcomes of the EMS based on Dec-RL : (a) the output power trajectories, (b) the FC modules profiles, and (c) the SoC changes.

Conclusions

A proof-of-concept study is performed to address the power allocation challenge of a heavy-duty HFCV in a decentralized fashion in this paper. Simulation and experimental results indicate that the suggested strategy with Dec-RL EMS can successfully allocate the FC modules' optimal powers and battery pack in the MFCV with different FCS modules parameters and operation preferences. Besides, the suggested data-driven approach is proved to outperform the Cen-RL algorithm in terms of convergence speed. The experiment results of the small-scale test bench substantiate that the decentralized EMS is implementable in real-time. The suggested decentralized RL method will be advanced through the deep learning method in our future work.

References

[1] Z. P. Cano, D. Banham, S. Ye, A. Hintennach, J. Lu, M. Fowler, *et al.*, "Batteries and fuel cells for emerging electric vehicle markets," *Nature Energy,* vol. 3, pp. 279-289, 2018/04/01 2018.

[2] H. S. Das, C. W. Tan, and A. H. M. Yatim, "Fuel cell hybrid electric vehicles: A review on power conditioning units and topologies," *Renewable and Sustainable Energy Reviews,* vol. 76, pp. 268-291, 2017/09/01/ 2017.

[3] T. Hua, R. Ahluwalia, L. Eudy, G. Singer, B. Jermer, N. Asselin-Miller*, et al.*, "Status of hydrogen fuel cell electric buses worldwide," *Journal of Power Sources,* vol. 269, pp. 975-993, 2014/12/10/ 2014.

[4] U. Eberle, B. Müller, and R. Helmolt, "Fuel cell electric vehicles and hydrogen infrastructure: Status 2012," *Energy & Environmental Science,* vol. 5, pp. 8790-8798, 07/30 2012.

[5] N. Marx, L. Boulon, F. Gustin, D. Hissel, and K. Agbossou, "A review of multi-stack and modular fuel cell systems: Interests, application areas and on-going research activities," *International Journal of Hydrogen Energy,* vol. 39, pp. 12101-12111, 2014.

[6] J. E. Garcia, D. F. Herrera, L. Boulon, P. Sicard, and A. Hernandez, "Power sharing for efficiency optimisation into a multi fuel cell system," in *2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)*, 2014, pp. 218-223.

[7] N. Marx, D. Hissel, F. Gustin, L. Boulon, and K. Agbossou, "On the sizing and energy management of an hybrid multistack fuel cell – Battery system for automotive applications," *International Journal of Hydrogen Energy,* vol. 42, pp. 1518-1526, 2017/01/12/ 2017.

[8] A. Khalatbarisoltani, J. C. O. Cepeda, L. Boulon, D. L. St-Pierre, J. Solano, and C. Duarte, "A New Real-Time Centralized Energy Management Strategy for Modular Electric Vehicles," in *2018 IEEE Vehicle Power and Propulsion Conference (VPPC)*, 2018, pp. 1-5.

[9] B. Somaiah and V. Agarwal, "Distributed Maximum Power Extraction From Fuel Cell Stack Arrays Using Dedicated Power Converters in Series and Parallel Configuration," *IEEE Transactions on Energy Conversion,* vol. 31, pp. 1442-1451, 2016.

[10] H. Zhang, X. Li, X. Liu, and J. Yan, "Enhancing fuel cell durability for fuel cell plug-in hybrid electric vehicles through strategic power management," *Applied Energy,* vol. 241, pp. 483-490, 2019/05/01/ 2019.

[11] A. O. M. Fernandez, M. Kandidayeni, L. Boulon, and H. Chaoui, "An Adaptive State Machine Based Energy Management Strategy for a Multi-Stack Fuel Cell Hybrid Electric Vehicle," *IEEE Transactions on Vehicular Technology,* pp. 1-1, 2019.

[12] T. Wang, Q. Li, H. Yang, L. Yin, X. Wang, Y. Qiu*, et al.*, "Adaptive current distribution method for parallel-connected PEMFC generation system considering performance consistency," *Energy Conversion and Management,* vol. 196, pp. 866-877, 2019/09/15/ 2019.

[13] A. Khalatbarisoltani, M. Kandidayeni, L. Boulon, and X. Hu, "Power Allocation Strategy based on Decentralized Convex Optimization in Modular Fuel Cell Systems for Vehicular Applications," *IEEE Transactions on Vehicular Technology,* pp. 1-1, 2020.

[14] U. DOE, "The fuel cell technologies office multi-year research, development, and demonstration plan," Technical report: US Department of Energy2016.

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

[16] B. Xu and X. Li, "A Q-learning based transient power optimization method for organic Rankine cycle waste heat recovery system in heavy duty diesel engine applications," *Applied Energy,* vol. 286, p. 116532, 2021/03/15/ 2021.

[17] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of real-world reinforcement learning," *arXiv preprint arXiv:1904.12901,* 2019.

[18] H. Zhang, J. Peng, H. Tan, H. Dong, and F. Ding, "A Deep Reinforcement Learning Based Energy Management Framework with Lagrangian Relaxation for Plug-in Hybrid Electric Vehicle," *IEEE Transactions on Transportation Electrification,* 2020.

[19] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 6059-6066.

[20] M. Zanon and S. Gros, "Safe reinforcement learning using robust MPC," *IEEE Transactions on Automatic Control,* 2020.

[21] L. Wen, J. Duan, S. E. Li, S. Xu, and H. Peng, "Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1-7.

[22] N. Jansen, B. Könighofer, S. Junges, A. Serban, and R. Bloem, "Safe Reinforcement Learning Using Probabilistic Shields," 2020.

[23] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[24] F. Berkenkamp, M. Turchetta, A. P. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *arXiv preprint arXiv:1705.08551,* 2017.

[25] X. Hu, T. Liu, X. Qi, and M. Barth, "Reinforcement Learning for Hybrid and Plug-In Hybrid Electric Vehicle Energy Management: Recent Advances and Prospects," *IEEE Industrial Electronics Magazine,* vol. 13, pp. 16-25, 2019.

[26] T. Liu, Y. Zou, D. Liu, and F. Sun, "Reinforcement learning of adaptive energy management with transition probability for a hybrid electric tracked vehicle," *IEEE Transactions on Industrial Electronics,* vol. 62, pp. 7837-7846, 2015.

[27] T. Liu, X. Hu, S. E. Li, and D. Cao, "Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle," *IEEE/ASME Transactions on Mechatronics,* vol. 22, pp. 1497-1507, 2017.

[28] Y. Zou, T. Liu, D. Liu, and F. Sun, "Reinforcement learning-based real-time energy management for a hybrid tracked vehicle," *Applied energy,* vol. 171, pp. 372-382, 2016.

[29] T. Liu, X. Hu, W. Hu, and Y. Zou, "A heuristic planning reinforcement learning-based energy management for power-split plug-in hybrid electric vehicles," *IEEE Transactions on Industrial Informatics,* vol. 15, pp. 6436-6445, 2019.

[30] G. Du, Y. Zou, X. Zhang, Z. Kong, J. Wu, and D. He, "Intelligent energy management for hybrid electric tracked vehicles using online reinforcement learning," *Applied Energy,* vol. 251, p. 113388, 2019.

[31] Q. Zhou, J. Li, B. Shuai, H. Williams, Y. He, Z. Li*, et al.*, "Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle," *Applied Energy,* vol. 255, p. 113755, 2019/12/01/ 2019.

[32] B. Shuai, Q. Zhou, J. Li, Y. He, Z. Li, H. Williams*, et al.*, "Heuristic action execution for energy efficient charge-sustaining control of connected hybrid vehicles with model-free double Q-learning," *Applied Energy,* vol. 267, p. 114900, 2020/06/01/ 2020.

[33] B. Xu, X. Hu, X. Tang, X. Lin, H. Li, D. Rathod*, et al.*, "Ensemble reinforcement learning-based supervisory control of hybrid electric vehicle for fuel economy improvement," *IEEE Transactions on Transportation Electrification,* vol. 6, pp. 717-727, 2020.

[34] B. Xu, X. Tang, X. Hu, X. Lin, H. Li, D. Rathod*, et al.*, "Q-Learning-Based Supervisory Control Adaptability Investigation for Hybrid Electric Vehicles," *IEEE Transactions on Intelligent Transportation Systems,* pp. 1-10, 2021.

[35] Y. Hu, W. Li, K. Xu, T. Zahid, F. Qin, and C. Li, "Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning," *Applied Sciences,* vol. 8, p. 187, 2018.

[36] Y. Wu, H. Tan, J. Peng, H. Zhang, and H. He, "Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus," *Applied Energy,* vol. 247, pp. 454-466, 2019/08/01/ 2019.

[37] X. Han, H. He, J. Wu, J. Peng, and Y. Li, "Energy management based on reinforcement learning with double deep Q-learning for a hybrid electric tracked vehicle," *Applied Energy,* vol. 254, p. 113708, 2019.

[38] X. Guo, T. Liu, B. Tang, X. Tang, J. Zhang, W. Tan*, et al.*, "Transfer deep reinforcement learning-enabled energy management strategy for hybrid tracked vehicle," *IEEE Access,* vol. 8, pp. 165837-165848, 2020.

[39] R. Lian, H. Tan, J. Peng, Q. Li, and Y. Wu, "Cross-type transfer for deep reinforcement learning based hybrid electric vehicle energy management," *IEEE Transactions on Vehicular Technology,* vol. 69, pp. 8367-8380, 2020.

[40] J. Zhou, S. Xue, Y. Xue, Y. Liao, J. Liu, and W. Zhao, "A novel energy management strategy of hybrid electric vehicle via an improved TD3 deep reinforcement learning," *Energy,* vol. 224, p. 120118, 2021.

[41] T. Liu, X. Tang, X. Hu, W. Tan, and J. Zhang, "Human-like energy management based on deep reinforcement learning and historical driving experiences," *arXiv preprint arXiv:2007.10126,* 2020.

[42] W.-S. Lin and C.-H. Zheng, "Energy management of a fuel cell/ultracapacitor hybrid power system using an adaptive optimal-control method," *Journal of Power Sources,* vol. 196, pp. 3280-3289, 2011/03/15/ 2011.

[43] R. C. Hsu, S. Chen, W. Chen, and C. Liu, "A Reinforcement Learning Based Dynamic Power Management for Fuel Cell Hybrid Electric Vehicle," in *2016 Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS)*, 2016, pp. 460-464.

[44] J. Yuan, L. Yang, and Q. Chen, "Intelligent energy management strategy based on hierarchical approximate global optimization for plug-in fuel cell hybrid electric vehicles," *International Journal of Hydrogen Energy,* vol. 43, pp. 8063-8078, 2018/04/19/ 2018.

[45] N. P. Reddy, D. Pasdeloup, M. K. Zadeh, and R. Skjetne, "An intelligent power and energy management system for fuel cell/battery hybrid electric vehicle using reinforcement learning," in *2019 IEEE Transportation Electrification Conference and Expo (ITEC)*, 2019, pp. 1-6.

[46] X. Lin, B. Zhou, and Y. Xia, "Online Recursive Power Management Strategy based on the Reinforcement Learning Algorithm with Cosine Similarity and a Forgetting Factor," *IEEE Transactions on Industrial Electronics,* pp. 1-1, 2020.

[47] H. Sun, Z. Fu, F. Tao, L. Zhu, and P. J. J. o. P. S. Si, "Data-driven reinforcement-learning-based hierarchical energy management strategy for fuel cell/battery/ultracapacitor hybrid electric vehicles," vol. 455, p. 227964, 2020.

[48] A. K. Soltani, M. Kandidayeni, L. Boulon, and D. L. St-Pierre, "Modular Energy Systems in Vehicular Applications," *Energy Procedia,* vol. 162, pp. 14-23, 2019/04/01/ 2019.

[49] T. Logenthiran, D. Srinivasan, and A. M. Khambadkone, "Multi-agent system for energy resource scheduling of integrated microgrids in a distributed system," *Electric Power Systems Research,* vol. 81, pp. 138-148, 2011.

[50] L. Busoniu, R. Babuska, 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, pp. 156-172, 2008.

[51] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *arXiv preprint arXiv:1911.10635,* 2019.

[52] J. Solano, S. Jemei, L. Boulon, L. Silva, D. Hissel, and M. C. Pera, "IEEE VTS Motor Vehicles Challenge 2020 - Energy Management of a Fuel Cell/Ultracapacitor/Lead-Acid Battery Hybrid Electric Vehicle," in *2019 IEEE Vehicle Power and Propulsion Conference (VPPC)*, 2019, pp. 1-6.

[53] J. S. Martinez, D. Hissel, M.-C. Péra, and M. Amiet, "Practical control structure and energy management of a testbed hybrid electric vehicle," *IEEE Transactions on Vehicular Technology,* vol. 60, pp. 4139-4152, 2011.

[54] M. Kandidayeni, A. M. Fernandez, L. Boulon, and S. Kelouwani, "Efficiency Upgrade of Hybrid Fuel Cell Vehicles' Energy Management Strategies by Online Systemic Management of Fuel Cell," *IEEE Transactions on Industrial Electronics,* pp. 1-1, 2020.

[55] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning,* vol. 8, pp. 279-292, 1992.

[56] H. Chen, P. Pei, and M. Song, "Lifetime prediction and the economic lifetime of Proton Exchange Membrane fuel cells," *Applied Energy,* vol. 142, pp. 154-163, 2015/03/15/ 2015.

[57] N. Herr, J.-M. Nicod, C. Varnier, L. Jardin, A. Sorrentino, D. Hissel*, et al.*, "Decision process to manage useful life of multi-stacks fuel cell systems under service constraint," *Renewable Energy,* vol. 105, pp. 590-600, 2017/05/01/ 2017.

[58] S. Satyapal, "U.S. Department of Energy Hydrogen and Fuel Cell Technology Overview " 2018.

[59] USDRIVE, "Fuel Cell Technical Team Roadmap," 2017.

[60] K. Mongird, V. V. Viswanathan, P. J. Balducci, M. J. E. Alam, V. Fotedar, V. S. Koritarov*, et al.*, "Energy Storage Technology and Cost Characterization Report," Pacific Northwest National Lab.(PNNL), Richland, WA (United States)2019.

[61] R. Liessner, J. Schmitt, A. Dietermann, and B. Bäker, "Hyperparameter Optimization for Deep Reinforcement Learning in Vehicle Energy Management," in *ICAART (2)*, 2019, pp. 134-144.

[62] F. C. Fernandez and W. Caarls, "Parameters tuning and optimization for reinforcement learning algorithms using evolutionary computing," in *2018 International Conference on Information Systems and Computer Science (INCISCOS)*, 2018, pp. 301-305.

[63] M. Kandidayeni, A. O. M. Fernandez, A. Khalatbarisoltani, L. Boulon, S. Kelouwani, and H. Chaoui, "An Online Energy Management Strategy for a Fuel Cell/Battery Vehicle Considering the Driving Pattern and Performance Drift Impacts," *IEEE Transactions on Vehicular Technology,* vol. 68, pp. 11427-11438, 2019.