Gestion de l'Energie basée sur l'Apprentissage par Renforcement pour Véhicule Hybride Pile à Combustible et Batterie

Antoine BÄUMLER[a]
Abdelmoudjib BENTERKI[a]
Jianwen MENG[a]
Toufik AZIB[a]
Mousa Boukhnifer[b]
ESTACA, ESTACA'Lab – Paris-Saclay F-78180 Montigny-le-Bretonneux, France[a]
Université de Lorraine, Cité Universitaire, Metz, 57000, Moselle, France[b]

31 mai 2023

Résumé
Notre société relève actuellement de nombreux défis liés à la transition énergétique, cela concerne particulièrement la nouvelle mobilité basée sur l'électrification. Dans cette étude, nous nous intéressons aux voies de développement et d'intégration du véhicule à hydrogène (FCHEV). Cette technologie s'appuyant le plus souvent sur un système hybride pile à combustible/Batterie, permet de reluire considérablement les émissions de polluants à l'usage tout en offrant des performances comparables aux solutions conventionnelles. Cela dépend du choix de l'architecture, le dimensionnement et surtout de la stratégie de gestion d'énergie embarquée. Pour ce qui est de la gestion d'énergie, de nombreuses techniques ont été proposées, et les plus récentes en date sont les techniques basées sur l'apprentissage, particulièrement le renforcement (Reinforcement Learning RL). Le RL permet un partage plus optimisé entre les sources en fonction des objectifs de consommation et de dégradation des composants, ainsi qu'une meilleure capacité à s'adapter à de nouvelles situations. L'approche proposée a été entraînée et validée en utilisant des cycles de conduite réels, dont un échantillon est présenté pour démontrer la faisabilité de l'approche.

1. INTRODUCTION

Dans un contexte de réduction des émissions de gaz à effet de serre (GES) et de polluants, ainsi que d'une raréfaction des matières premières, le besoin de trouver de nouvelles sources d'énergie pour la mobilité est grandissant. En effet, actuellement, dans le domaine du transport, en particulier dans le secteur routier, l'électrification est largement adoptée [1]. La technologie ayant actuellement le vent en poupe est le véhicule à batterie avec la technologie lithium-ion mais cela requiert une utilisation massive de métaux rares une autonomie limitée et un temps de recharge relativement long. Une autre solution pour l'électrification est l'utilisation de piles à combustible (PAC), fonctionnant de préférence à l'hydrogène pour son absence d'émission de polluants à l'utilisation. Les véhicules hybrides hydrogène-batterie permettent de réduire la quantité de métaux rares nécessaire, d'avoir un temps de recharge comparable aux véhicules thermiques, tout en n'émettant pas de GES et de polluants à l'utilisation. En revanche le mode de production de l'hydrogène doit encore être questionné, avec la majeure partie de la production (>90%) d'origine fossile et donc très émettrice de GES [2]. Les principaux défis des véhicules fonctionnant avec une PAC sont d'être compétitif avec les véhicules avec moteur à combustion internes et les batteries. Notamment sur les problématiques de puissance de sortie, d'autonomie et de temps de recharge. En revanche la durée de vie du système est encore un défi à relever. Un des levier pour améliorer la compétitivité des véhicules hybrides électriques à pile a combustible (FCHEV) est l'amélioration des algorithmes de gestion de l'énergie. Les premières stratégies étaient basées sur les règles, c'est-à-dire que la gestion se fait par la définition de règles préétablies par des experts. Les techniques employées pouvaient être thermostatiques [3], basées sur la logique floue ou la stratégie de suivi de la puissance [4]. Les techniques basées sur les règles reposent sur les connaissances des experts en termes de situations de conduites rencontrées, cela ne permettant pas d'atteindre un fonctionnement idéal sur les objectifs fixés. Pour cela, une nouvelle famille de méthodes s'est imposée : les stratégies d'optimisation. Le principe de fonctionnement étant de définir une fonction d'objectif qui intègre les indicateurs à considérer, comme la consommation et la dégradation. Elles se séparent en deux sous-familles, les optimisations globales et locales. Les optimisations globales cherchent à atteindre les points de fonctionnements menant à une optimisation globale de la fonction d'objectif. La méthode la plus utilisée ici est la programmation dynamique [5], dont le principe est de réduire un problème en plusieurs petits sous-problèmes d'optimisation permettant d'atteindre un mode de fonctionnement optimal sur tout le cycle. Théoriquement, si le pas de discrétisation est infiniment petit, la fonction d'objectif atteint son minimum. Mais cela requiert une connaissance précise sur tout le cycle de conduite, ce qui n'est rarement le cas dans la gestion d'énergie et de plus la puissance de calcul nécessaire augmente de manière exponentielle avec la taille du pas de discrétisation. La programmation dynamique est donc en général plutôt utilisée comme méthode de comparaison pour les autres EMS [6]. Les méthodes d'optimisation locale, cherchent à chaque instant de chercher le point de fonctionnement optimal, ce qui est plus réaliste en termes d'implémentation pour le fonctionnement temps réel car l'optimisation ne dépend que des états précédents et le temps de calcul est considérablement réduit. Les méthodes que l'on peut citer sont le principe de minimisation de Pontryagin's (PMP) [7] et la stratégie de minimisation de consommation d'énergie (ECMS) [8]. Ces méthodes d'optimisation permettent de mieux réduire la consommation et même la dégradation en comparaison des méthodes basées sur les règles, mais l'optimisation est toujours lourde en calcul et s'adapte peu à des cycles de conduite plus complexes qui n'ont pas été pris en compte dans le paramétrage

de l'EMS. Pour cela, une nouvelle famille d'EMS émerge, plutôt basée sur l'apprentissage, avec notamment l'apprentissage par renforcement (RL). Cela requiert moins de puissance de calcul et une meilleure adaptabilité aux nouveaux cycles de conduite [9].
Le fonctionnement détaillé du RL est détaillé en section 3.2. La contribution se présente par une nouvelle approche dans la conception de la fonction d'objectif, appelée de fonction de récompense en RL, et dans le type d'action utilisé en sortie de l'EMS.
Dans la littérature, les fonctions de récompense se composent d'une première partie pour optimiser la consommation, d'une seconde sur la dégradation et la dernière, nécessaire au fonctionnement même de l'EMS, la gestion de l'état de charge (SOC) de la batterie. Cette dernière partie est nécessaire, car elle a pour rôle d'inciter l'EMS à conserver un certain niveau de charge, autrement la batterie serait systématiquement déchargée due à l'incitation à minimiser l'utilisation de la PAC par les autres composantes de la fonction de récompense. La méthode usuellement employée pour la gestion du SOC est le maintien de charge, qui consiste à imposer un SOC de référence [6] [10]. Une amélioration, en termes de vitesse d'apprentissage, a été proposée par [11], où l'agent est récompensé lorsque la variation du SOC tend vers la référence. Une autre étude [12] se démarque avec une gestion de SOC se basant en partie sur les limites de SOC.
La fonction de récompense est basée sur le coût opérationnel de fonctionnement de chaque composant, permettant indirectement de pénaliser l'utilisation de la batterie. La principale contribution ici est d'intégrer des limites de SOC plutôt que d'imposer un SOC de référence. Cela implique un certain défi au niveau de la convergence de l'algorithme de RL, en particulier pour les méthodes récentes avec les actions continues. Cela a le potentiel d'offrir une plus grande liberté pour l'EMS pour l'optimisation des autres composantes de la fonction de récompense, notamment la consommation et la dégradation.
L'article est décomposé avec une section 2 présentant la modélisation du FCHEV avec les détails sur la PAC et la batterie, ensuite dans la section 3 la problématique de gestion d'énergie est détaillé et le principe de l'apprentissage par renforcement est expliqué

2. MODÉLISATION DU SYSTÈME

Le système considéré ici est une pile à combustible connectée avec un convertisseur DC-DC unidirectionnel et une batterie avec un convertisseur DC-DC bidirectionnel, permettant sa recharge comme présenté en Figure.1. La modélisation des convertisseurs a été omise en raison de sa grande demande en puissance de calcul. La puissance demandée est générée à partir

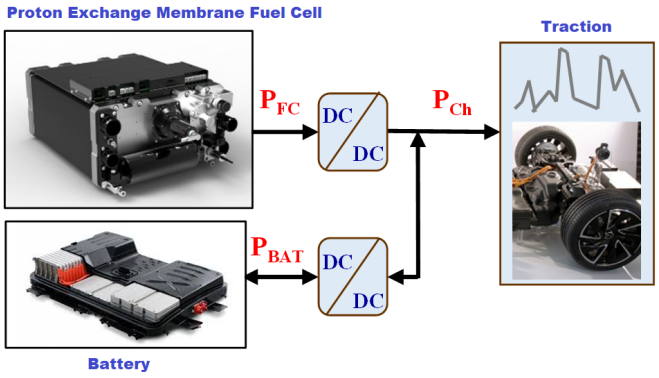

FIG. 1. Architecture du véhicule hybride PAC/batterie

de profils de vitesse venant de conditions de conduites réelles, auxquelles le principe fondamental de la dynamique est appliqué.

2.1. *PEMFC model*

La PAC est modélisé par un modèle statique et son comportement est caractérisé par sa courbe de polarisation en Figure.2 et en suivant l'équation qui suit :

$$V_{FC}(i) = n_{cells} \cdot (V_0 - A \cdot ln(\frac{i}{i_0}) - r \cdot i + ln(1 - \frac{i}{i_l})) \quad (1)$$

avec n_{cells} le nombre de cellules dans la pile ; A, r et B sont respectivement les coefficients empiriques des pertes par activation, ohmique et par concentration, i_0 est la densité courant minimum et i_l la densité de courant maximum. Additionnelle-

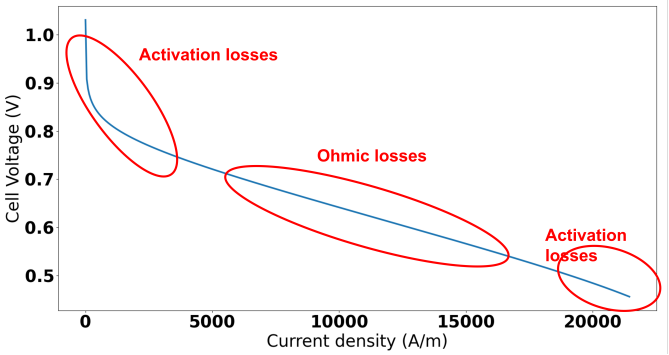

FIG. 2. FC courbe de polarisation

ment, la consommation de carburant ou d'hydrogène est proportionnelle au courant de la PAC

$$\dot{m}_{H_2} = \frac{n_{cells} \cdot M_{H_2}}{2F} \cdot i_{fc} \quad (2)$$

Avec M_{H_2} la masse molaire du dihydrogène, F est la constante Faraday, et i_{fc} le courant de la PAC

2.2. *Battery model*

La batterie est modélisée à l'aide d'un circuit double RC comme montré dans la Figure.3 Ce qui se modélise par les équa-

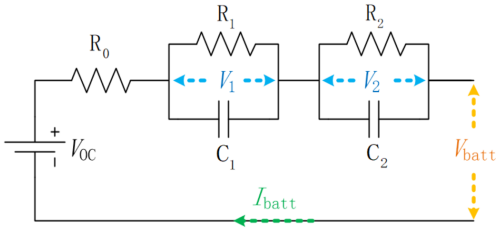

FIG. 3. Double RC circuit Battery model [13]

tions suivantes :

$$V_{cell}(t) = V_{OC}(SOC) - V_1(t) - V_2(t) - R_0 \cdot I_{bat}(t) \quad (3)$$

Avec le SOC, l'état de charge de la batterie. La mise à jour des variables d'états V_1, V_2 et SOC est discrétisée avec la méthode des bloqueurs d'ordre zéro détaillé dans [14].

3. GESTION DE L'ÉNERGIE (EMS)

3.1. *Problématique de l'EMS*

Le premier objectif de l'EMS est de respecter les contraintes des composants qui sont les limites de puissance de sortie des sources et leurs dynamiques. Et, avec la présence d'un moyen de stockage d'énergie, il est nécessaire d'avoir un moyen de gérer intelligemment l'état de charge. Ensuite, l'objectif est aussi d'optimiser la consommation, et de réduire la dégradation des

composants. Ici seul la consommation est considérée, afin de montrer l'efficacité de la nouvelle méthode proposée pour la gestion du SOC. En effet, dans la littérature, la méthode largement utilisée est la stratégie de maintien de charge, dont le principe est d'imposer un SOC de référence, et l'EMS est incitée à conserver un SOC proche de cette valeur. Cette stratégie a pour conséquence de limiter l'utilisation complète de la batterie avec un SOC qui va rester uniquement autour de la valeur de référence. Initialement, cette méthode est là pour inciter l'EMS à charger la batterie et de la faire fonctionner dans un espace optimal pour la batterie en termes de dégradation et de performances. Autrement, si dans la fonction de coût, seulement la consommation est prise en compte, l'EMS aura tendance à ne jamais démarrer la PAC, amenant à une décharge complète de la batterie, mettant le véhicule à l'arrêt, car aucune énergie ne sera amenée aux moteurs. Une amélioration dans l'approche serait d'imposer des limites de SOC plutôt qu'une référence. Cela consiste à donner une récompense constante sur la plage de charge tolérée, et lorsque que les limites sont atteintes, l'EMS est pénalisée progressivement à mesure qu'elle s'éloigne des limites. Les limites seraient de la forme suivante :

$$L_{bounds} = \begin{cases} tanh(w_{SOC} \cdot (SOC_{min} - SOC)) \\ \quad \text{if } SOC < SOC_{min} \\ tanh(w_{SOC} \cdot (SOC - SOC_{max})) \\ \quad \text{if } SOC > SOC_{max} \\ 0 \quad \text{else} \end{cases} \tag{4}$$

où L_{bounds} est le coût associé aux limites de SOC, w_{SOC} est le poids associé au SOC. La fonction $tanh$ est ici pour garder le coût entre les valeurs de -1 et 1.

3.2. *L'apprentissage par renforcement*

Le RL est composé de deux entités, un agent, prenant les décisions et l'environnement, retournant l'état du système, et un retour d'expérience par le biais récompense associé. L'interaction entre les deux entités est représentée en Figure.4. L'environnement doit suivre le principe des chaînes Markovienne pour que l'apprentissage puisse converger, cela veut dire qu'il faut qu'à chaque étape, l'état renvoyé doit contenir toutes les informations nécessaires à la compréhension de l'état de l'environnement. Dans le cas étudié, en gestion d'énergie, l'agent serait l'EMS et l'environnement, le véhicule avec les sources d'énergie et son comportement. Le but du RL est, à chaque pas, d'estimer la valeur de chaque couple état-action, et, dans les cas d'une politique gourmande (greedy policy), l'action choisie sera celle avec la valeur la plus élevée. Dans le cas du Q-learning, ces valeurs sont stockées dans un tableau. La valeur en question est la récompense attendue et est mise à jour selon l'équation suivante [15] :

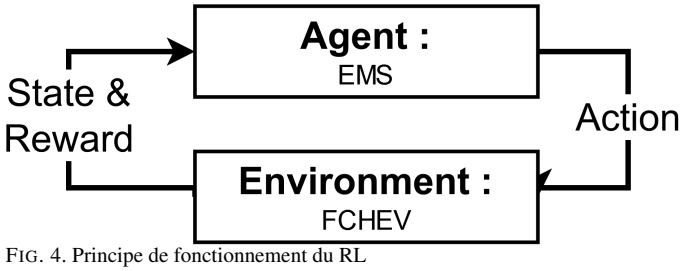

FIG. 4. Principe de fonctionnement du RL

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \cdot (R_t + \gamma \cdot \underset{a}{max}(Q(s_{t+1}, a)) - Q(s_t, a_t)) \tag{5}$$

Avec $Q(s_t, a_t)$ le tableau des couples état-action à l'état et action 't', α le taux d'apprentissage dont la valeur doit être un compromis entre vitesse d'apprentissage et précision, R_t la récompense obtenue à l'instant t, γ le coefficient de réduction, paramètre réglant à quel point l'agent va s'intéresser aux récompenses futures. Sa valeur est usuellement 0.99, et lorsqu'elle est nulle, l'agent cherchera à obtenir la meilleure récompense instantanée sans se soucier des récompenses futures. En revanche, cette technique ne prend en entrée que des états discrets et ne donne que des actions discrètes, rendant la précision et la performance du système très limitée et est sujet au problème de dimensionnalité. Depuis, des nouvelles méthodes ont émergé, et ont permis de rendre les états et ensuite les actions continues. Le Deep Q-learning permet d'évaluer les couple états-actions à l'aide d'un réseau de neurones. Mais les actions sont toujours discrètes, rendant la prise de décision plus lente et moins précise. Récemment, des méthodes avec des états et action continues ont émergé, telles que le Deep Deterministic Policy Gradient (DDPG) [16]. Le principe de fonctionnement est d'avoir deux réseaux de neurones, appelés l'acteur et le critique. Le rôle de l'acteur est de choisir les actions et le critique d'évaluer l'action de l'agent selon un état donné. Des techniques plus avancées, comme le Soft Actor Critic (SAC), ont un réseau de neurones supplémentaire, estimant uniquement la valeur de l'état. Ces réseaux de neurones additionnels aident à la convergence de l'acteur vers une politique optimale pour la fonction de récompense donnée. Cela mène à une partie importante dans la conception du modèle d'apprentissage, le choix du type d'action pour la gestion de l'énergie, les informations liées à l'état à fournir à l'agent et surtout la fonction de récompense, avec la nouvelle approche proposée.

3.3. *L'apprentissage par renforcement appliqué à la gestion d'énergie*

Dans la littérature, il y a deux approches pour le type d'action pour l'agent, soit la demande de puissance directe [6] [11], soit la variation de puissance [12] [17] de la PAC. La méthode de variation de puissance permet plus facilement de contraindre l'agent à ne pas dépasser les limites de dynamique de la PAC, voire en limiter les dégradations si suffisamment contraintes. Dans cet article, la stratégie adoptée est d'avoir comme action la variation de densité courant, méthode d'autant plus simple à l'application au regard du modèle utilisé. En effet, le modèle de PAC donné dans la section.2.1 la valeur en entrée est une demande en courant. De plus, les convertisseurs eux même fonctionnent sur une demande en courant, facilitant donc le fonctionnement global du système. Cela se matérialise par :

$$Action = \{\Delta i\} \tag{6}$$

$$i_{dem}(t+1) = i(t) + \Delta i \tag{7}$$

Avec i la densité de courant, i_{dem} la densité de courant demandés à la PAC.
En ce qui concerne les états, du système, les variables choisies dans la littérature sont généralement les mêmes avec le SOC, la puissance demandée par la charge, et dans le cas des actions avec la variation de puissance, la puissance actuelle de la PAC. Ce qui donne :

$$Etat = [SOC(t), P_{load}(t), P_{fc}(t)] \tag{8}$$

Avec $P_{load}(t)$ la puissance demandée par le groupe motopropulseur, $P_{fc}(t)$ la puissance de sortie de la PAC à l'instant t et la vitesse du véhicule. Ces variables donnent toutes les informations nécessaires pour la prise de décision pour une EMS.
Et pour la fonction de récompense, comme détaillée précédemment dans la problématique sous-section 3.1, prend deux composantes, les limites d'états de charges et la consommation de la PAC :

$$R_{bounds} = L_{bounds} - w_{conso} \cdot \dot{m}_{fc} \tag{9}$$

Avec w_{conso} le poids associé à la consommation de la PAC. Dans le choix des limites de SOC, il était nécessaire d'ajouter une pénalité augmentant de manière linéaire, en raison de l'aspect éparpillé de la récompense. Cela a en fait pour effet de

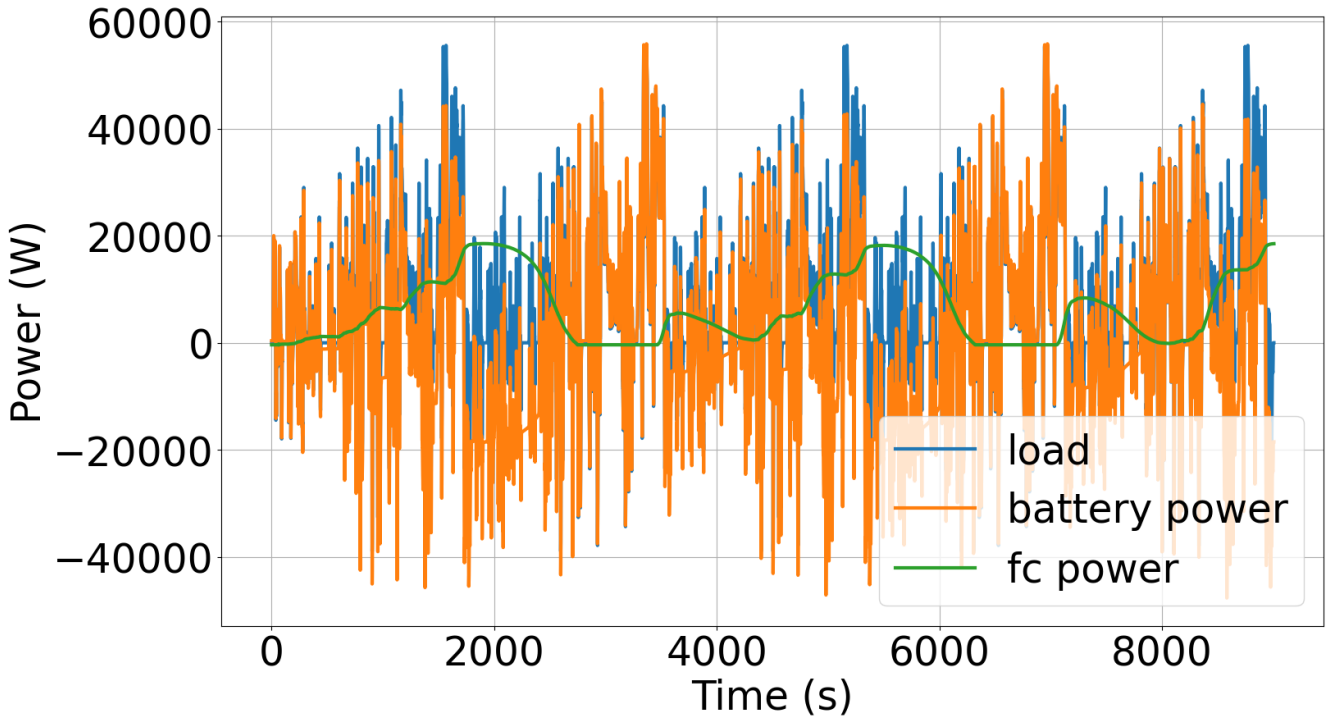

FIG. 5. Partage de la puissance demandé entre les deux sources

rendre l'apprentissage plus complexe comme cela a été détaillé dans [18]. Et malgré le lissage, l'apprentissage reste long et peu stable.

4. RÉSULTATS

Le modèle et l'algorithme d'apprentissage fonctionnent sous Python v3.10. L'agent est entraîné sur 69 cycles de conduite basée sur des enregistrements de conditions réelles. Un cycle de la même base de données est utilisé pour la validation, dans le but d'observer les tendances d'apprentissage. Pour les tests et l'évaluation des résultats, une répétition de cinq cycles de conduite WLTP est employée. La méthode proposée est comparée aux autres rencontrées dans la littérature. Les fonctions de récompenses retenues pour la comparaison sont :

$$R = -w_1 \cdot \dot{m}_{H_2}(t) - w_2 \cdot (SOC(t) - SOC_{ref})^2 \qquad (10)$$

La gestion d'énergie est montrée en Figure.5 avec la puissance demandée et la puissance des deux sources, PAC et batterie. La puissance demandée est bien respectée, en revanche, on remarque une forte dynamique dans la puissance de la batterie, pouvant dégrader ses performances sur le long terme. Cette dégradation peut être réduite en changeant le dimensionnement de la batterie, avec une batterie ayant, certes, moins de capacité, mais plus tolérante aux dynamiques de la puissance. Mais en ce qui concerne la puissance de la PAC, les variations sont lentes et les puissances employées sont faibles, ce qui est optimal pour les dégradations. Mais régulièrement la PAC est éteinte et redémarrée, un grand facteur dans la dégradation. De plus, les puissances de la PAC sont proches du fonctionnement optimal en termes de consommation avec un fonctionnement proche du rendement optimal des systèmes PAC. Dans la Figure.6, la stratégie du maintien de charge a tendance à faire varier la puissance de PAC plus fréquemment, ce qui est moins optimal en termes de dégradation. Ce comportement est dû à l'incitation par l'objectif de conserver une valeur de SOC proche de la référence. Figure.7 présente l'évolution du SOC durant le cycle de conduite.

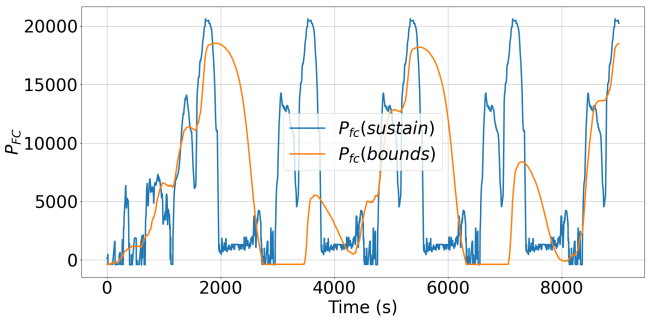

FIG. 6. Puissance de PAC pour chaque stratégie

Les variations de SOC pour la stratégie proposée sont bien plus notables pour les limites de SOC. En effet, la stratégie basée sur des limites donne plus de liberté dans l'utilisation de la batterie pour l'EMS. Cela permet donc de faciliter l'optimisation sur les autres composantes de la fonction de récompense.

Dans la Figure.8, est affichée l'évolution de la récompense pour chaque stratégie. La fonction de récompense proposée converge plus vite vers une solution que celle avec le maintien de charge.

Dans le tableau 1 est indiqué la consommation des deux méthodes. La méthode proposée consomme donc 5.4% d'hydrogène en moins, validant la performance de l'approche.

Méthode	**Consommation de H_2(kg)**
Limite SOC	1.086673
Maintien de charge	1.148293

TABLEAU 1. Consommation de chacune de stratégies

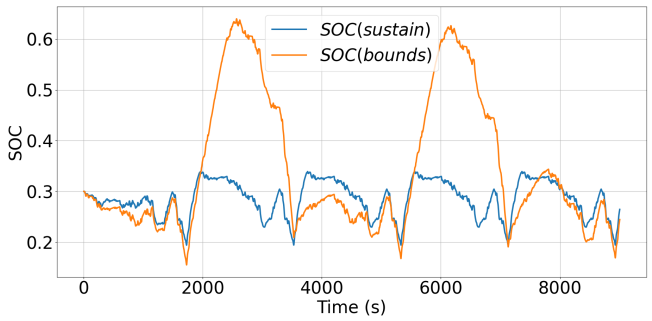

FIG. 7. Évolution du SOC pour chaque stratégie

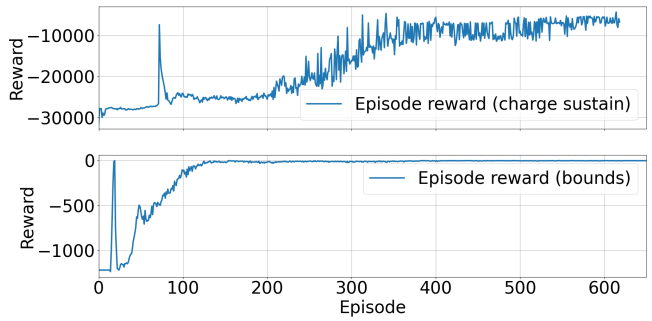

FIG. 8. Évolution de l'apprentissage pour chaque stratégie

5. CONCLUSION

Cette étude propose une EMS basée sur l'apprentissage, avec la méthode SAC et une nouvelle approche dans la fonction de récompense. L'efficacité de cette dernière a été démontrée en comparaison des méthodes précédente basées sur le maintien de charge avec une consommation réduite de 5.4%. De plus, même la vitesse de convergence avec la fonction de récompense proposée est plus grande et la stabilité est meilleure. En revanche, l'EMS de cette étude, démarrent et éteignent fréquemment la PAC, ce qui cause un vieillissement accéléré. Pour cela, des recherches plus approfondies doivent être menées afin d'éviter ce genre de comportement. Une solution simple qui pourrait être mise en avant est de pénaliser par le biais de la fonction de récompense l'arrêt de la PAC, ou selon certaines conditions, d'imposer une puissance minimale à la PAC pour l'empêcher de s'éteindre.

6. RÉFÉRENCES

[1] Yujie Wang, Li Wang, Mince Li, and Zonghai Chen. A review of key issues for control and management in battery and ultra-capacitor hybrid energy storage systems. *eTransportation*, 4 :100064, May 2020.

[2] U.P.M. Ashik, W.M.A. Wan Daud, and Hazzim F. Abbas. Production of greenhouse gas free hydrogen by thermocatalytic decomposition of methane – A review. *Renewable and Sustainable Energy Reviews*, 44 :221–256, April 2015.

[3] Ding Gen Li and Dai Wei Feng. Thermostatic Control for Series Hydraulic Hybrid Vehicle (SHHV) Energy Management. *Advanced Materials Research*, 512-515 :2676–2681, May 2012.

[4] F. X. Chen, Y. Yu, and J. X. Chen. Control System Design of Power Tracking for PEM Fuel Cell Automotive Application. *Fuel Cells*, 17(5) :671–681, October 2017.

[5] Shiyong Tao, Weirong Chen, Rui Gan, Luoyi Li, Guorui Zhang, Ying Han, and Qi Li. Energy management strategy based on dynamic programming with durability extension for fuel cell hybrid tramway. *Railway Engineering Science*, 29(3) :299–313, September 2021.

[6] Renzong Lian, Jiankun Peng, Yuankai Wu, Huachun Tan, and Hailong Zhang. Rule-interposing deep reinforcement learning based energy management strategy for power-split hybrid electric vehicle. *Energy*, 197 :117297, April 2020.

[7] K. Ettihir, L. Boulon, and K. Agbossou. Optimization-based energy management strategy for a fuel cell/battery hybrid power system. *Applied Energy*, 163 :142–153, February 2016.

[8] Jie Li, Yonggang Liu, Datong Qin, Guang Li, and Zheng Chen. Research on Equivalent Factor Boundary of Equivalent Consumption Minimization Strategy for PHEVs. *IEEE Transactions on Vehicular Technology*, 69(6) :6011–6024, June 2020.

[9] Teng Teng, Xin Zhang, Han Dong, and Qicheng Xue. A comprehensive review of energy management optimization strategies for fuel cell passenger vehicle. *International Journal of Hydrogen Energy*, 45(39) :20293–20303, August 2020.

[10] Li Tang, Giorgio Rizzoni, and Simona Onori. Energy Management Strategy for HEVs Including Battery Life Optimization. *IEEE Transactions on Transportation Electrification*, 1(3) :211–222, October 2015. Number : 3.

[11] Kai Deng, Yingxu Liu, Di Hai, Hujun Peng, Lars Löwenstein, Stefan Pischinger, and Kay Hameyer. Deep reinforcement learning based energy management strategy of fuel cell hybrid railway vehicles considering fuel cell aging. *Energy Conversion and Management*, 251 :115030, January 2022.

[12] Peng Wu, Julius Partridge, and Richard Bucknall. Cost-effective reinforcement learning energy management for plug-in hybrid fuel cell and battery ships. *Applied Energy*, 275 :115258, October 2020.

[13] Jianwen Meng. Diagnostic de la batterie et gestion de l'énergie pour applications embarquées.

[14] Jianwen Meng, Meiling Yue, and Demba Diallo. Nonlinear extension of battery constrained predictive charging control with transmission of Jacobian matrix. *International Journal of Electrical Power & Energy Systems*, 146 :108762, March 2023.

[15] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8 :279–292, 1992.

[16] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, July 2019. arXiv :1509.02971 [cs, stat].

[17] Lei Deng, Shen Li, Xiaolin Tang, Kai Yang, and Xianke Lin. Battery thermal- and cabin comfort-aware collaborative energy management for plug-in fuel cell electric vehicles based on the soft actor-critic algorithm. *Energy Conversion and Management*, 283 :116889, May 2023.

[18] Guillaume Matheron, Nicolas Perrin, and Olivier Sigaud. The problem with DDPG : understanding failures in deterministic environments with sparse rewards. volume 12397, pages 308–320. 2020. arXiv :1911.11679 [cs, stat].