Energy Conversion and Management 292 (2023) 117362

Contents lists available at ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

Health-considered energy management strategy for fuel cell hybrid electric vehicle based on improved soft actor critic algorithm adopted with Beta policy

Weiqi Chen [a], Jiankun Peng [a,*], Jun Chen [a], Jiaxuan Zhou [a], Zhongbao Wei [b], Chunye Ma [a]

[a] School of Transportation, Southeast University, Nanjing 211102, China
[b] School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Keywords:
Fuel cell hybrid electric vehicle
State of health
Soft actor critic
Beta policy
Multi-objective optimization

ABSTRACT

Deep reinforcement learning-based energy management strategy (EMS) is essential for fuel cell hybrid electric vehicles to reduce hydrogen consumption, improve health performance and maintain charge. This is a complex nonlinear constrained optimization problem. In order to solve the problem of high bias caused by the inconsistency between the infinite support of stochastic policy and the bounded physics constraints of application scenarios, this paper proposes the Beta policy to improve standard soft actor critic (SAC) algorithm. This work takes hydrogen consumption, health degradation of both fuel cell system and power battery, and charge margin into consideration to design an EMS based on the improved SAC algorithm. Specifically, an appropriate tradeoff between money cost during driving and charge margin is firstly determined. Then, optimization performance differences between the Beta policy and the standard Gaussian policy are presented. Thirdly, ablation experiments of health constraints are conducted to show the validity of health management. Finally, comparison experiments indicate that, the proposed strategy has a 5.12% performance gap with dynamic programming-based EMS with respect to money cost, but is 4.72% better regarding to equivalent hydrogen consumption. Moreover, similar performances in validation cycle demonstrate good adaptability of the proposed EMS.

1. Introduction

The traditional transportation sector contributes approximately 20% of global greenhouse gas emissions and air pollution [1], which is a heavy burden on environment protection and energy security. Automobile companies and research institutes have been working to develop new vehicles to replace conventional internal combustion engine vehicles. At present, there are three mainstream technologies [2]: hybrid electric vehicles (HEV), fuel cell electric vehicles and pure electric vehicles. The HEV are only transitional product because they cannot avoid the high emissions and pollution of the internal combustion engine. Although pure electric vehicles have no emissions and pollution, they suffer from short driving range and long charging time due to the limitation of power battery technology [3].

In recent years, fuel cells have attracted more and more attentions because of high efficiency, no pollution, fast-fueling and low noise [4,5]. However, fuel cells have the disadvantages of slow dynamic response and poor stability in fast power demand conditions [6]. In order to ensure the sustainability of output power, a power battery with high energy density is generally equipped to be used together with fuel cells as an auxiliary energy source [7]. The power battery pack provides peak power to smooth fluctuations of fuel cells' output power [8]. However, hybrid energy storage makes the power and energy flow of the vehicle more complicated, so it is of great significance to formulate efficient and reasonable energy management and optimization strategies, so as to give full play to the performance and advantages of fuel cell hybrid electric vehicles (FCHEV).

As a fundamental and key technology of FCHEV, energy management strategy (EMS) is designed to distribute energy between fuel cell system (FCS) and power battery to improve powertrain efficiency and fuel economy [9], and also reduce health degradation of fuel cell and power battery [10]. The reported EMS can be classified as three categories [11]: 1) rule-based, 2) optimization-based, 3) learning-based.

The rule-based EMSs have the advantages of simplicity, easy implementation, low computation burden, and good reproducibility [12]. However, the rules are highly dependent on engineering

* Corresponding author.
E-mail address: jkpeng@seu.edu.cn (J. Peng).

https://doi.org/10.1016/j.enconman.2023.117362
Received 5 May 2023; Received in revised form 29 June 2023; Accepted 30 June 2023
Available online 5 July 2023

experiences and intuition of automobile engineers, and the optimization performance are not ideal [13]. In order to achieve optimal control performance, more and more researchers have developed optimization-based EMSs, which can be divided into off-line global optimization EMS and on-line instantaneous optimization EMS [14]. Dynamic programming (DP) is the most classical global optimal control algorithm, and is used as comparison benchmarks. Ali [15] designed a FCHEV state space for DP-based EMS, then conducted over different driving cycles by using an emulation test-rig. His simulation results revealed an improvement in energy efficiency up to 29% compared to rule-based EMS. In order to improve the durability of fuel cells and power battery while reducing hydrogen consumption, Liu [16] developed a multi-objective EMS for FCHEV based on the non-dominated sorting genetic algorithm (NSGA). Although the off-line global optimization-based EMS can achieve global optimum performance, the dependence on prior knowledge of drive cycles and heavy computation burden makes them not feasible for real-time implementation [17].

With the research of instantaneous optimization methods, abundant on-line optimization-based EMSs have been developed by scholars [18]. Based on the Pontryagin's Minimum Principle (PMP), the constrained global optimization problem can be simplified to a Hamiltonian local minimization problem, thus reducing computation burden and providing real-time application capability [19]. Jiang developed a real-time EMS for FCHEV based on the two-dimension PMP. And simulation results indicated that his proposed EMS can approximate the optimal performance of DP-based EMS [20]. In order to better measure performances of various components in hybrid systems, the equivalent consumption minimization strategy (ECMS) is derived by introducing an equivalent factor to represent equivalent energy consumption of auxiliary power sources [21]. Although the ECMS can be easily utilized as instantaneous optimization-based EMS, the performances are seriously influenced by the equivalent coefficient [22]. To compensate for this shortcoming, adaptive equivalent consumption minimization strategy (AECMS) has been developed. Li proposed EMS for FCHEV based on AECMS by adjusting equivalent factor according to health state of power sources, to make sure the charge sustenance of battery and prolong the lifetime of fuel cell [23]. Furthermore, the model predictive control (MPC) algorithm, which is based on feedback correction and rolling optimization, has also been introduced to solve the constrained nonlinear dynamic energy management problem. Guo designed a fuel cell engine friendly real-time EMS in signal intersection scenario, where the MPC-based EMS can reduce the equivalent hydrogen consumption by approximately 3.04% [24]. Although the EMSs based on instantaneous optimization methods have relatively fewer computation burden and easier implementation, which make them available in real-time applications. However, compared to the global optimization-based EMSs, there is still great room for performance improvement of local instantaneous optimization-based EMSs.

Thanks to the development of artificial intelligence technology, EMSs based on reinforcement learning (RL) and deep reinforcement learning (DRL) algorithms have been widely studied in recent years. The unexpected model sensitivity of optimization-based methods can be eliminated by RL algorithms, whose architecture are model-free [25]. As the basic RL methods, Q-learning [26,27] and Dyna [28] were firstly employed to optimize EMS, but the discretization of state and action made them not favorable for solving high-dimension optimization problems. Thus, DRL methods quickly took RL's place in energy management and optimization field. Based on the deep Q-learning (DQN) method, Tang proposed a longevity-conscious EMS for FCHEV. His simulation performances reached 88.73 % of the DP-based EMS in standard driving cycle but with only 30% computation burden [11]. However, the DQN uses the same values both to choose and evaluate actions, making it easily to fall into local optimum which is caused by over optimistic value estimates [29]. In this regard, Han proposed a double deep Q-learning (DDQN)-based EMS structure, and achieved 7.1% improvement in terms of fuel economy than the DQN-based EMS [30]. However, the DQN series methods face the same dilemma, that is, the discrete action space may lead to dimension curse [31]. Thus, deep deterministic policy gradient (DDPG) algorithm with continuous state and action space was introduced into EMS to optimize problems with high-dimension action space [32]. Li took history cumulative trip information into state vector of DDPG algorithm to develop EMS for HEV, and achieved an average 3.5% gap from DP benchmark [33]. Wu combined Gumbel-Softmax action noise with the twin delayed deep deterministic policy gradient (TD3) algorithm, and designed EMS for a plug-in HEV. His experiment results revealed only 2.55% gap from the DP benchmark in terms of fuel economy [34]. In order to solve the problem of high sensitivity of hyperparameter and insufficient exploration of strategy space, Wu developed soft actor critic (SAC)-based EMS to optimize power allocation of HEV and enhance battery thermal safety and health performance [35]. In our previous work, a simple application of SAC algorithm in the field of FCHEV energy management is presented [36]. DRL-based EMSs can be trained on cloud servers and transmitted to vehicle terminals through the V2X (vehicle to everything) communication network for real-time application [37]. More importantly, they can achieve near global optimum performance.

As an advanced DRL algorithm, SAC has shown better convergence and lower sensitivity of hyper parameters than others [38]. SAC is based on the maximum entropy DRL framework, in which the actor aims to simultaneously maximize expected reward and entropy to enhance exploration. Different from afore mentioned ones, the SAC algorithm outputs Gaussian distribution for action choosing in stochastic continuous control problems. In many application scenarios, the action spaces are bounded due to physics constraints, which means actor can only take actions within a finite interval. However, the Gaussian distribution is infinite support, which will unavoidably introduce an estimation bias due to boundary effects [39]. This will slow down training process and even lead to worse convergence and performance. To avoid the unexpected high bias, an adoption of stochastic distribution with finite support boundary is necessary for SAC algorithm. Considering that the Beta distribution is a continuous probability distribution defined in $[0, 1]$ interval, its continuity and boundedness perfectly avoid the negative influences caused by the Gaussian distribution [40].

This paper aims to bridge the afore mentioned research gaps, and proposes an energy management strategy for FCHEV based on the improved SAC algorithm, where the health performances of both fuel cell system and power battery are considered. The main contributions are as follows.

(1) An intelligent energy management framework based on the improved SAC algorithm is proposed for FCHEV, considering health degradation of both fuel cell system and power battery pack.

(2) Given the high estimation bias caused by the inconsistency between the infinite support of Gaussian policy and the bounded physics constraints of vehicle powertrain, the Beta distribution is adopted to improve optimization performance of SAC-based EMS for the first time.

(3) An appropriate weight coefficient is determined after numerous simulation experiments. And ablation experiments for health performance are emphasized to validate the effectiveness of the proposed EMS in reducing driving cost and prolong lifetime of FCHEV.

(4) The proposed SAC-based EMS is better than other DRL-based EMSs and achieves very close performance to the DP benchmark. And simulation results in different driving cycles demonstrate its excellent adaptability.

The remainder of the article is organized as follows. The vehicle powertrain system, fuel cell system, power battery model, and their health models are described in Section 2. The details of SAC algorithm and training setup are presented in Section 3. Simulation results are discussed in Section 4. Section 5 concludes this paper.

2. Model description

2.1. Powertrain structure

A fuel cell hybrid electric bus is selected as research object in this paper, driven by an electric motor with 200 kW. Main configuration of FCHEV is listed in Table 1. As shown in Fig. 1, power is supplied by a fuel cell stack and a power battery. The requested traction force of the FCHEV is calculated based on the longitudinal dynamics as follows:

$$F_t = mgfcos\theta + mgsin\theta + \frac{AC_D v_h^2}{21.15} + \delta m a_h \# \tag{1}$$

where m is total mass of the vehicle, f is the coefficient of rolling resistance, θ is the road slope which is considered to be 0 in this study, A is the frontal area, C_D is the coefficient of air resistance, δ is the coefficient of rotation mass convention, and g denotes the acceleration of gravity. Then, rotation speed of wheel W_w and torque of the drive shaft T_w are expressed as follows:

$$\begin{cases} W_w = v_h / r \\ T_w = r_w \bullet F_t \end{cases} \# \tag{2}$$

where r_w is wheel radius. The rotation speed W_m and torque T_m of motor are then calculated as follows:

$$\begin{cases} W_m = W_w \bullet R_{fd} \\ T_m = \begin{cases} T_w / (\eta_{fd} \bullet R_{fd}), T_w \geq 0 \\ T_w \bullet \eta_{fd} / R_{fd}, T_w < 0 \end{cases} \# \end{cases} \tag{3}$$

where R_{fd} is the final drive gear ration, and η_{fd} is the efficiency of drive shaft. The power requested by vehicle is calculated as follows:

$$P_{req} = \begin{cases} T_m \bullet W_m / \eta_m, T_m \geq 0 \\ T_m \bullet W_m \bullet \eta_m, T_m < 0 \end{cases} \# \tag{4}$$

where η_m is the efficiency of motor, interpolated from efficiency map based on the quasi-steady motor model, as illustrated in Fig. 2**(b)**. The FCHEV with a fuel cell system and a pack of Li-ion power battery. Correspondingly, the P_{req} can be expressed as follows:

$$P_{req} = P_{DC/DC} + P_{bat} \# \tag{5}$$

where $P_{DC/DC}$ is the output power of DC/DC converter, which is determined by the output power of fuel cell system (P_{fcs}) according to experimental fitting data. And P_{bat} is power of Li-ion power battery pack, including charge and discharge. Since the total power (P_{req}) is determined, the value of P_{bat} is indirectly determined by P_{fcs}.

2.2. Fuel cell model

The fuel cell system including a fuel cell stack and auxiliary components, serving as the primary source of power for FCHEV, electrochemically converts the chemical energy of hydrogen and oxygen into electrical energy. Physical and empirical model are utilized to simulate the fuel cell system in this work, by considering physical laws and operating conditions. The hydrogen consumption rate of fuel cell stack $\dot{m}$ can be calculated [41]:

$$\dot{m} = \frac{P_{fcs}}{\eta_{fcs} \bullet L_v} \# \tag{6}$$

where L_v represents the hydrogen lower heating value, equaling to $120kJ/g$, and η_{FCS} represents the efficiency of fuel cell stack. The relationships between the output power of fuel cell system P_{fcs} and hydrogen consumption rate $\dot{m}$ and efficiency η_{fcs} are illustrated in Fig. 2 **(a)**.

Four different types of disadvantageous driving conditions—load changing cycle, start-stop cycle, low-power load, and high-power load—are the primary contributors of fuel cell degradation. According to the findings of Song et.al [42], total performance degradation of fuel cell system $D_{fcs}(\%)$ can be formulated by the discrete expression:

$$D_{fcs} = \sum_{t=0}^{n} \left[d_{ss}(t) + d_{low}(t) + d_{high}(t) + d_{cha}(t) \right] \# \tag{7}$$

where n is the number of time steps, $d_{ss}(t)$, $d_{low}(t)$, $d_{high}(t)$, $d_{cha}(t)$ are the health degradation caused by start-stop cycle, low-power load, high-power load, and load changing cycle at time step t respectively. More accurate calculation methods can be found in reference [43].

2.3. Power battery model

As the other energy source of FCHEV, the power battery system leverages Li-ion Battery pack to provide peak power and store extra power. As shown in Fig. 3**(b)**, the power battery system is simulated and modelled by a coupled electro-thermal-aging model comprising three sub-models: a second-order RC electro model, a two-state thermal model and an energy-throughput aging model [35]. In the second-order RC electro model, two RC branches are utilized to simulate the polarization effects. Then, the governing equations are given by [44]:

$$\frac{dSoC(t)}{dt} = \frac{I(t)}{3600 C_n} \# \tag{8}$$

$$\frac{dV_{p1}(t)}{dt} = -\frac{V_{p1}(t)}{R_{p1}(t) C_{p1}(t)} + \frac{I(t)}{C_{p1}(t)} \# \tag{9}$$

$$\frac{dV_{p2}(t)}{dt} = -\frac{V_{p2}(t)}{R_{p2}(t) C_{p2}(t)} + \frac{I(t)}{C_{p2}(t)} \# \tag{10}$$

$$V_t(t) = V_{oc}(SoC) + V_{p1}(t) + V_{p2}(t) + R_s I(t) \# \tag{11}$$

where $I(t)$ and $V_t(t)$ are the load current and terminal voltage at time step t, V_{p1} and V_{p2} are the polarization voltage across the RC branches which are parameterized by the capacitance C_{p1}, C_{p2} and resistance R_{p1}, R_{p2}. Based on the thermal energy conservation principle, the following equations are given:

$$C_c \frac{dT_c(t)}{dt} = \frac{T_s(t) - T_c(t)}{R_c} + H(t) \# \tag{12}$$

$$C_s \frac{dT_s(t)}{dt} = \frac{T_c(t) - T_s(t)}{R_c} + \frac{T_f(t) - T_s(t)}{R_u} \# \tag{13}$$

$$T_a(t) = \frac{T_c(t) + T_s(t)}{2} \# \tag{14}$$

where T_s, T_c, T_a and $T_f(t)$ are temperature of battery surface, core, internal average and ambient respectively, all in the unit of ℃. R_c and R_u are the thermal resistances owing to the heat conduction inside the battery and the convection at battery surface. C_c and C_s are equivalent thermal capacitance of the battery core and surface. Heat generation

Table 1
Main configuration of the FCHEV.

Items	Parameters	Value
Vehicle	Curb weight	8400 kg
	Tire radius	0.46 m
	Front windward area	6.56 m^2
	Final reducer ratio	6.2
	Rolling resistance coefficient	0.012
	Air resistance coefficient	0.55
Motor	Peak power	200 kW
	Efficiency	[0.77, 0.98]
FCS	Peak power	60 kW
DC-DC converter	Peak power	60 kW
	Efficiency	[0.90, 0.95]
Battery	Capacity	108.14 kWh

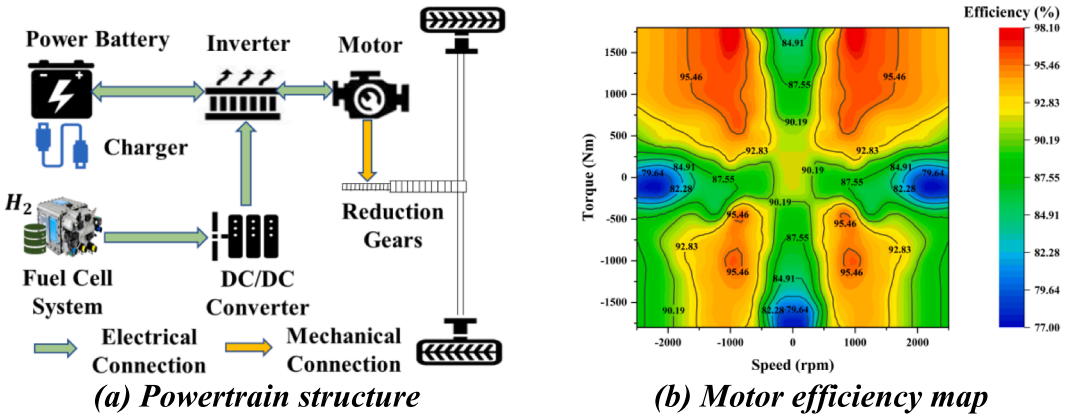

Fig. 1. Powertrain structure and motor efficiency map of the FCHEV.

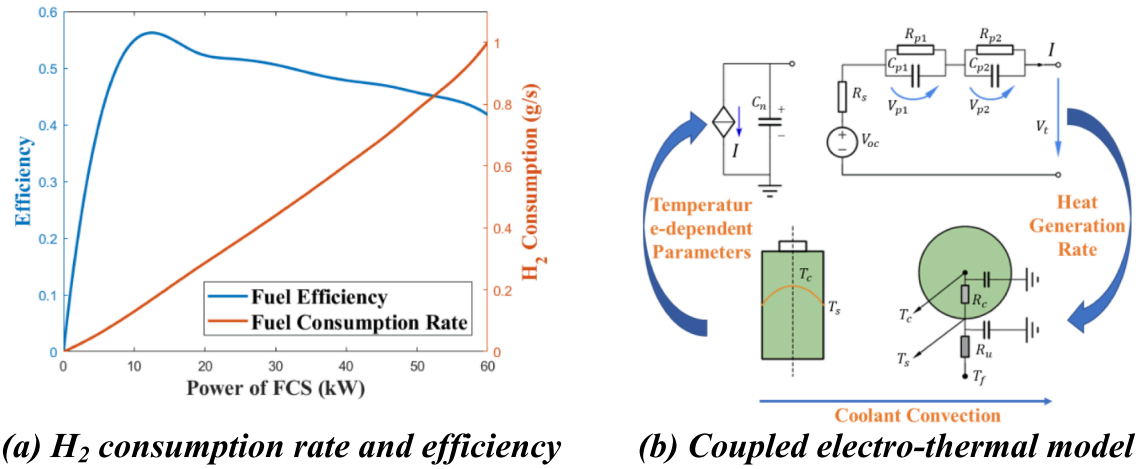

Fig. 2. Fuel cell system output characteristics and power battery model.

rate $H(t)$ inside battery system is subject to ohmic heat, polarization heat and irreversible entropy heat together. Therefore, the heat generation rate can be calculated as follows:

$$H(t) = I(t)\left[V_{p1}(t) + V_{p2}(t) + R_s(t)I(t)\right] + I(t)[T_a(t) + 273]E_n(SoC, t) \tag{15}$$

where E_n represents the entropy change during electrochemical reactions. The energy-throughput model is adapted to evaluate battery degradation, presuming that the battery can sustain a specific amount of accumulated charge flow before being scrapped [45]. Therefore, the dynamic of SOH (state of health) is calculated by:

$$\Delta SOH_t = -\frac{|I(t)|\Delta t}{2N(c, T_a)C_n}\# \tag{16}$$

where Δt is the current duration. $N(c, T_a)$ is the equivalent number of cycles till the battery system reaches its end of life. Based on Arrhenius equation, the impact of C-rate (c) and internal temperature are taken into account. The percentage of capacity loss ΔC_n then can be expressed as follows:

$$\Delta C_n = B(c) \bullet \exp\left(-\frac{E_a(c)}{RT_a}\right) \bullet Ah^z\# \tag{17}$$

where $B(c)$ denotes pre-exponential factor which is fitted by experimental data, R is the ideal gas constant which equals to 8.314 $J/(mol \bullet K)$, z is the power-law factor equals to 0.55, Ah represents the ampere-hour throughput, and E_a denotes activation energy in the unit of J/mol defined by [44]:

$$E_a(c) = 31700 - 370.3 \bullet c\# \tag{18}$$

The battery reaches its end of life when the C_n drops by 20%. According to this definition and referring to [46], Ah and N can be expressed as follows:

$$Ah(c, T_a) = \left[\frac{20}{B(c)} \bullet \exp\left(-\frac{E_a(c)}{RT_a}\right)\right]^{1/z}\# \tag{19}$$

$$N(c, T_a) = 3600 \bullet Ah(c, T_a)/C_n\# \tag{20}$$

Finally, the health degradation condition of power battery pack for given current, temperature, and service dynamics can be explored according to **Equation (16)**.

3. SAC-based EMS

3.1. Standard SAC algorithm

Unlike conventional DRL algorithms, such as DQN and DDPG, which only focus on how to maximize the cumulative discounting rewards, the SAC algorithm aims to maximize expected reward while also maximizing policy entropy [47]. It is based on the actor-critic framework, where the actor outputs stochastic policy to enhance exploration of potential optimal actions. The critic is defined by the state-action value function, noted as Q and parameterized by θ, which is formulated by soft Bellman iteration:

$$Q(s_t, a_t) = r_t + \gamma E_{s_{t+1}, a_{t+1}}[Q(s_{t+1}, a_{t+1}) - \alpha \log(\pi(a_{t+1}|s_{t+1}))]\# \tag{21}$$

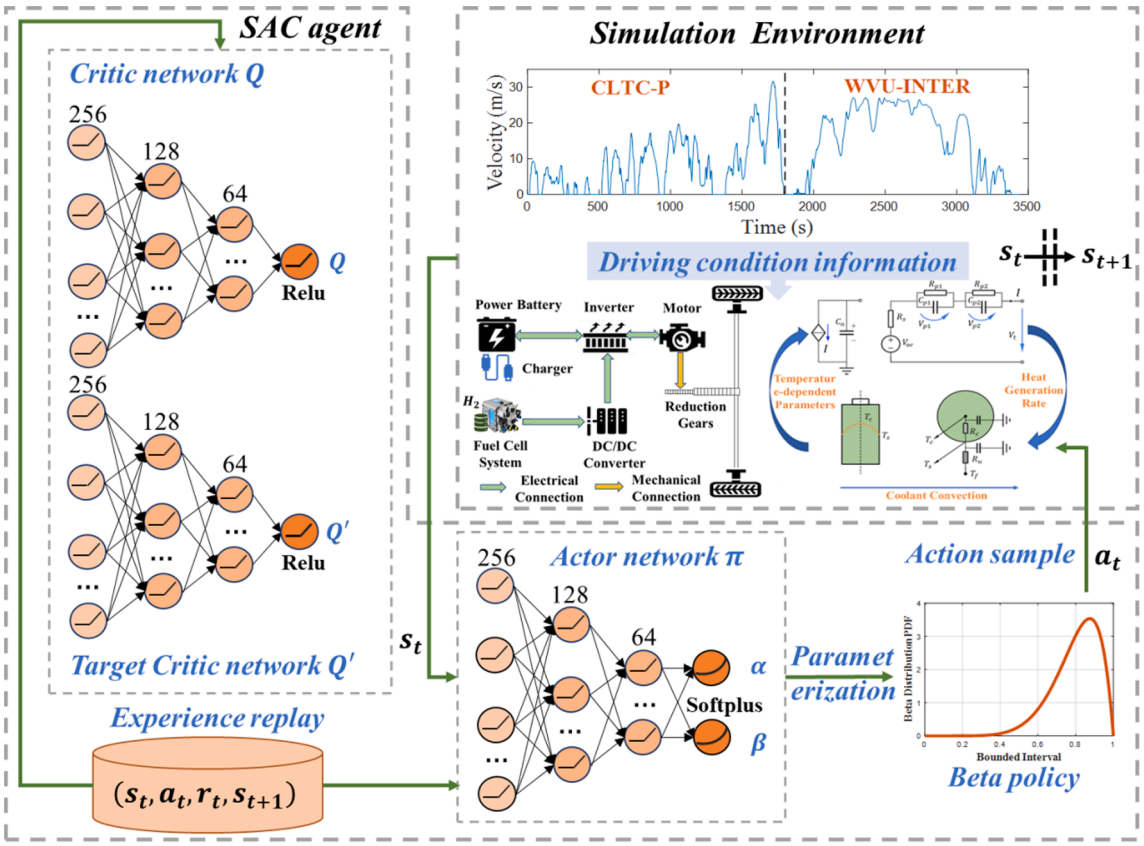

Fig. 3. The overall architecture of the proposed SAC-EMS.

where s_t, a_t, r_t are state, action, reward with respect to step t respectively, and s_{t+1}, a_{t+1} are state and action after state transition. γ is the discount factor, and E denotes mathematical expectation. α is the temperature factor to adjust the relative importance of the entropy term versus the reward, tuning automatically through neural network. π denotes the policy function and is parameterized by ϕ.

The critic networks can be trained by minimizing the soft Bellman residual:

$$J_Q(\theta) = E_{(s_t,a_t,r_t,s_{t+1})\ M}$$
$$\frac{1}{2}[Q(s_t,a_t) - [r_t + \gamma(Q^{'}(s_{t+1},a_{t+1}) - \alpha\log(\pi(a_{t+1}|s_{t+1}))\,)\,]]^2\# \tag{22}$$

where M is experience replay pool, and (s_t, a_t, r_t, s_{t+1}) are minibatches sampled from it randomly. The target critic network with parameter $[03B8]'$ is built to stabilize and speed up training process, whose soft update is controlled by the step factor τ:

$$\theta' \leftarrow (1-\tau)\theta' + \tau\theta\# \tag{23}$$

As for the policy function $\pi(a_t|s_t)$, for each state, it can be improved according to the information projection defined in terms of the Kullback-Leibler divergence:

$$\pi = \text{argmin}D_{KL}\left[\pi_\phi(\bullet|s_t)\|\frac{\exp(Q(s_t,\bullet)/\alpha)}{Z(s_t)}\right]\# \tag{24}$$

where $D_{KL}(\bullet)$ is the KL divergence. $Z(s_t)$ is logarithm partition function normalizing the distribution, and it does not contribute to gradient with respect to the new policy.

Then, the parameters of policy function can be learned by directly minimizing the expected KL divergence:

$$J_\pi(\phi) = E_{s_t\ M}[E_{a_t\ \pi}[\alpha\log(\pi(a_t|s_t)\,) - Q(s_t,a_t)\,]]\# \tag{25}$$

At each timestep, action is determined by current policy, which is reparametrized by following fixed distributions, such as Gaussian distribution.

$$a_t = f_\phi^\mu(s_t) + \varepsilon_t \odot f_\phi^\sigma(s_t)\# \tag{26}$$

where ε_t is an input noise vector. $f_\phi(s_t)$ is the output of policy network, which is divided into f_ϕ^μ that output the mean value, and f_ϕ^σ that output standard deviation of action distribution, so called Gaussian policy.

Temperature factor α is regulated automatically, its gradients is computed with the following objective:

$$J(\alpha) = E_{a_t\ \pi_\phi}\left[-\alpha log\pi_\phi(a_t|s_t) - \alpha\overline{H}\right]\# \tag{27}$$

where target entropy $\overline{H}$ is the opposite number of action dimension.

3.2. Beta policy

Firstly, the Gaussian policy of standard SAC algorithm is defined as follows [48]:

$$\pi_\phi(x|s) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\# \tag{28}$$

where $\mu = f_\phi^\mu(s)$ and $\sigma = f_\phi^\sigma(s)$ are mean value and standard deviation of normal distribution respectively, and they are output from the policy neural network π_ϕ. Then, the action can be determined by **Equation (26)**.

The action space of EMS is finite, while the Gaussian policy corresponds to an infinite support probability distribution, thus introducing bias. In order to fully explore strategy space in early stage of training, a larger σ value is required, but this will result in larger bias. Moreover, the action outputted by Gaussian policy can be executed by DRL agent only after truncation operation. The truncated action is also used for

computing the state value function and logarithmic probability gradient. Not only does it suffer from the same bias problem, another bias is also introduced through the subtraction of the baseline function.

In order to eliminate the influence of bias on algorithm performance, a policy with finite support probability distribution is needed. Thus, we introduce the Beta policy, according to the definition of Beta distribution, it can be presented as follows [39]:

$$\pi_\phi(x|s) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \# \tag{29}$$

where α and β are the shape parameters of the Beta distribution, and they are output from policy neural network with parameter ϕ, $\alpha = f_\phi^\alpha(s)$, $\beta = f_\phi^\beta(s)$. While $\Gamma(n) = (n-1)!$ is Gamma function that extends factorial to real number. The most significant difference between Beta policy and Gaussian policy is that, the Beta distribution has a bounded interval, as shown in Fig. 4. And it describes the probability of success, where $\alpha-1$ and $\beta-1$ can be thought of as the counts of successes and failures. The Beta policy is bias free, due to no probability density falls outside the boundary. And we only consider $\alpha, \beta > 1$, corresponding the case that the Beta distribution is concave and unimodal.

3.3. State and action space and reward function

The state space is a set of observations of the DRL agent on the interactive environment. It consists of the key variables that play significant roles in energy management decision, which is defined as: $[SOC, SOH_{bat}, SOH_{fcs}, P_{bat}, P_{fcs}, v, a]$. Since the state vector is sent into networks for calculation after normalization, their value bounds need to be determined, and the details are listed in Table 2. Because vehicle is stationary at the beginning, the initial states are $[0.5, 1.0, 1.0, 0, 0, 0, 0]$. The action space formulates what the DRL agent executes. According to powertrain topology of the FCHEV, action is defined as the output power of fuel cell system, P_{fcs}.

In our work, the energy management strategy has three optimization objectives: 1) reduce hydrogen consumption from fuel cell system; 2) reduce health degradation of both power battery and fuel cell system; 3) keep SOC of power battery in reasonable margin. It is generally agreed that, fuel consumption is a short-term issue, while health degradation represents long-term planning. But in fact, SOH degradation of power components continuously occurs during vehicle operation. And the one-time replacement cost can be decomposed into SOH cost at each time step. Thus, according to the equivalent cost minimization principle, a unified measurement of health degradation and fuel consumption is achieved through money cost. Finally, the reward function is derived as follows.

$$r_t = -\left[\rho_1 \dot{m}(t) + \rho_2 D_{fcs}(t) + \rho_3 \Delta SOH(t) + \omega |SOC(t) - SOC_{ref}|\right] \# \tag{30}$$

where ρ_1, ρ_3, ρ_3 are hydrogen price, replacement price of fuel cell system and power battery pack respectively. Thus, the former two objectives, fuel consumption and health degradation, are measured uniformly by money cost of driving. The weight coefficient ω determines the relative importance of the money cost versus battery SOC value, and should be fully explored in order to obtain better optimization performance. SOC_{ref} is the reference value of SOC, 0.5, the same as its initial value in this work.

Table 2
The value boundary of state variables.

Variable	Value boundary	According to
SOC	$[0, 1]$	Physical definition
SOH_{bat}	$[0, 1]$	Physical definition
SOH_{fcs}	$[0, 1]$	Physical definition
P_{bat}	$[-200, 200]kW$	Vehicle characteristics
P_{fcs}	$[0, 60]kW$	Vehicle characteristics
v	$[0, 120]km/h$	Road speed limit
a	$[-3, 2]m/s^2$	Vehicle characteristics

3.4. Training setup

In order to enable the proposed energy management strategy can cope with different traffic flow conditions, a driving cycle which covers low-speed to medium-speed and high-speed is constructed, named *Mix-train*, as shown in Fig. 5**(a)**. It consists of the China light-duty vehicle test cycle-passenger car (CLTC-P) and West Virginia University Interstate (WVU-INTER) cycles. To verify that the learned strategy is not trapped in overfitting, a driving cycle which consists of totally different driving conditions is constructed, named *Mix-valid*, as shown in Fig. 5**(b)**. It covers city and freeway scenarios, which consists of West Virginia University City (WVU-CITY) and Highway Fuel Economy Test (HWFET). The *Mix-valid* cycle is used to validate the adaptability of the learned policy after the training process is finished. The travel distance of the *Mix-train* cycle is 39.438 km, while the *Mix-valid* cycle is 21.822 km. The main hyper parameters of neural networks are presented in Table 3.

4. Simulation results and discussions

In this section, five main simulation results are discussed. Firstly, an appropriate value of weight coefficient ω, is determined after full comparative experiments. Secondly, the performance differences between Gaussian policy and Beta policy are presented. Thirdly, ablation experiments demonstrate the validity of health management. Then, horizontal comparison experiments show the excellent performance of the proposed algorithm. Last but not least, evaluation results in the prebuilt cycle illustrate pretty adaptability of the learned strategy.

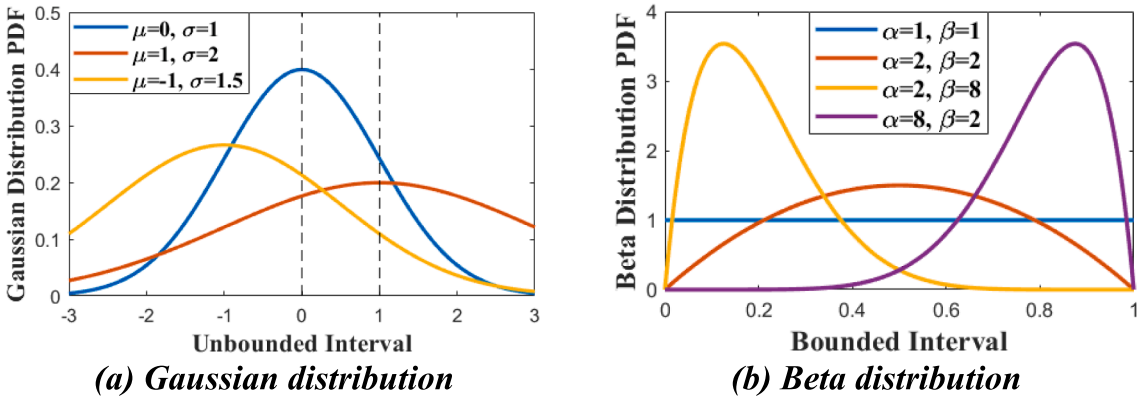

Fig. 4. Probability Density Function diagram.

(a) Mix-train, training set

(b) Mix-valid, validation set

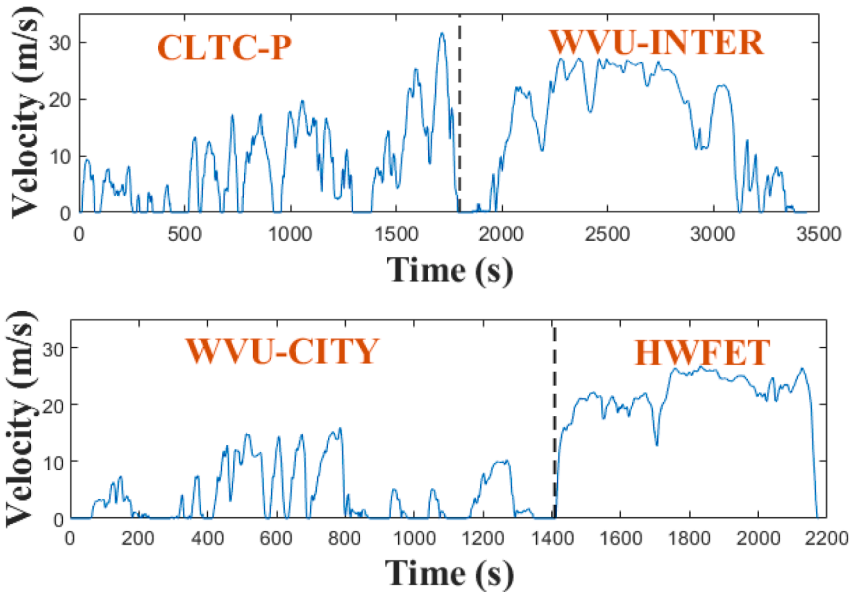

Fig. 5. The mixed driving cycles.

Table 3
Key hyper parameters of the DRL-EMS.

Parameter Description	Value
Number of hidden layers of *Actor* and *Critic* (target) network	2
Neurons distribution of *Actor* and *Critic* (target) network	256, 256
Neural network connection	Fully connected
Optimizer	Adam
Learning rate scheduler	Cyclical [49]
Learning rate range of *Actor*	[5e-5, 5e-3]
Learning rate range of *Critic*	[5e-5, 5e-4]
Discount factor	0.99
Step factor	0.001
Size of experience replay buffer	4e4
Minibatch size	64
Number of training episodes	400

4.1. Multiple objectives tradeoff

There are three objectives of energy management in this work: 1) decrease hydrogen consumption; 2) reduce SOH degradation of both fuel cells and power battery; 3) maintain SOC within reasonable range. As stated in **Section 3.3**, the hydrogen consumption and the SOH degradation can be normalized as money cost during driving. Thus, the weight coefficient ω, determining the relative importance of SOC target versus the money cost, is required to keep consistent magnitude order due to its immeasurability. Therefore, appropriate values of ω need to be fully investigated in order to achieve better performances. In this section, optimization performances of different ω values are analyzed from these aspects: the episode average reward, money spent per 100 km, equivalent hydrogen consumption per 100 km, FCS operating efficiency, SOH end value, and SOC trajectory.

Firstly, Fig. 6**(a)** illustrates the episode average reward when different $\omega = \{1, 100, 200, 300, 400, 500\}$ are implemented in the training process. It can be observed that all the tests have shown good convergence except $\omega = 1$. The curves of $\omega = 100$ and $\omega = 200$ converge at faster rates than others. Meanwhile, curve of $\omega = 100$ presents the most stable convergence and the highest reward among all the coefficients.

Secondly, the money cost and the equivalent hydrogen consumption are investigated. As displayed in Fig. 6**(b)**, the $\omega = 100$ curve dominates most of the lowest part of the curves distribution for the money comparisons, while obtaining the lowest end value, reflecting excellent economy. Meanwhile, Fig. 6**(c)** illustrates that the $\omega = 100$ curve dips to the lowest equivalent hydrogen consumption, while maintaining relatively fast convergence rate and continuous descending trend in subsequent training episode. More importantly, the lowest fuel consumption is clearly achieved when $\omega = 100$.

Thirdly, as illustrated in Fig. 6**(d)**, it is apparent that for $\omega = 100$, the fuel cell system operates in the low efficiency interval [0, 0.3] by the third lowest proportion. Furthermore, $\omega = 100$ maintains the upper proportion distribution in the moderate efficiency intervals [0.3, 0.4], [0.4, 0.45] and [0.45, 0.5]. While in the high efficiency intervals, [0.5, 0.55] and [0.55, 0.6], $\omega = 100$ shows the highest proportion. Overall, the fuel cell system operates more frequently in high-efficiency intervals when $\omega = 100$, so it can be concluded that the fuel cell system performs the superior efficiency when $\omega = 100$.

Last but not least, focused on the health performance and SOC trajectory. As displayed in Fig. 6**(e)**, the highest SOH end value of fuel cell system is achieved when $\omega = 100$, representing the best health management. Although the SOH of power battery is not very good. As the SOC trajectories shown in Fig. 6**(f)**, all curves remain within the [0, 1] interval, but several of them are very close to 0 during the [2500, 3000] time period. Running vehicle under very low SOC conditions will reduce the operation efficiency of power battery. But $\omega = 100$ effectively avoided this situation and achieved the highest SOC end value. The above indicate that the proposed strategy of $\omega = 100$ not only maintains the SOC in reasonable range but also explores health performance in considerable depth.

In summary, the proposed strategy successfully strikes an effective trade-off between the money cost and SOC target. At the same time, $\omega = 100$ presents the superior performances compared with other weight coefficients during training. Hence, $\omega = 100$ will be implemented in subsequent experiment phases.

4.2. Analysis of Beta policy

As shown in Fig. 7**(a)**, the Beta policy has much better convergence performance than the Gaussian policy. The Beta policy converges at 84th episode with a convergence reward of −15.879, while the Gaussian policy converges at 224th episode with a convergence reward of –23.353. The convergence time of Beta policy is only 37.5% that of Gaussian policy, but the convergence value is 47% higher. Better convergence performance comes from better action choices. The

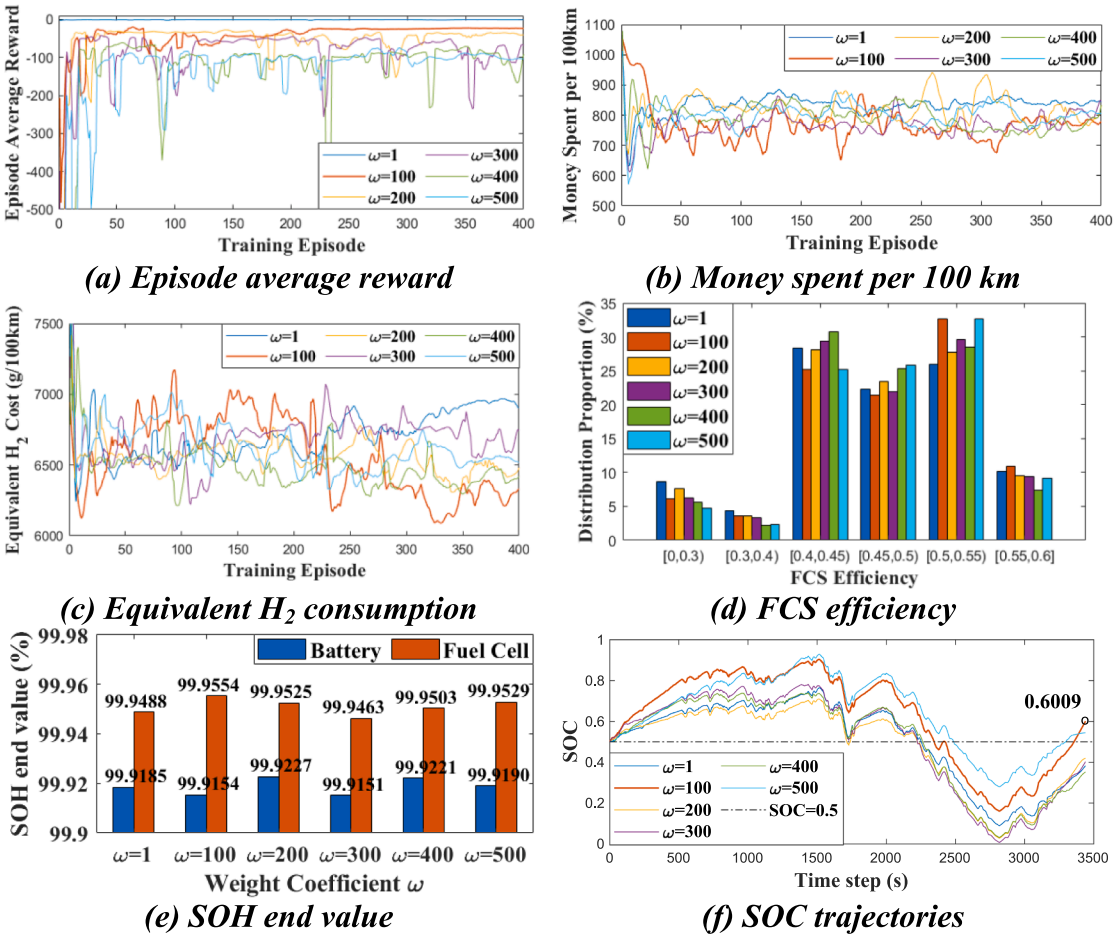

Fig. 6. Comparison of performances under different coefficients..ω

severity factor map of power battery with different policies is illustrated in Fig. 7**(b)**, which directly indicates the severity of power battery health degradation caused by the actions performed. It can be observed that compared with the Beta policy, the Gaussian policy causes the power battery to operate more frequently in high temperature and current rate areas, which accelerates battery health degrade and deteriorates energy management performance. The fuel cell working efficiency distribution proportion is presented In Fig. 7**(c)**. It can be seen, that the operation points of the Beta policy are much more distributed in the high-efficiency range. This will undoubtedly reduce hydrogen consumption and fuel cell health degradation, and thus improve the performance of fuel cell system. Fig. 7**(d)** illustrates the SOH trajectories of fuel cells and power battery with different policies. It can be seen that the Beta policy has much better health performance than that of Gaussian policy, which just confirms the above results. In summary, due to the finite support coped with physics constraints, the Beta policy can choose more reasonable actions than the Gaussian policy does, and thus obtains better optimization performance.

4.3. Validity of health constraint

In this section, ablation experiments for health management are conducted, and the comparison of different methods is shown in Table 4. The SOH trajectories of power battery with different baselines are shown in Fig. 8**(a)**, and the proposed method presents the best health performance. While Baseline B outperforms Baseline A and C, which means that adding a battery SOH-related item to the reward function can effectively inhibit battery degradation. The same phenome can also be observed in Fig. 8**(b)**, in which Baseline A is better than Baseline B and C while close to the proposed strategy.

Working in high C-rate and high temperature are two primary factors causing power battery degradation [35]. Fig. 8**(c)** illustrates battery severity factor maps of the proposed strategy and Baseline A. It can be observed that, compared to Baseline A, the proposed strategy maintains higher frequency in low C-rate and medium temperature areas. While Baseline A often works in high-temperature areas close to 60 degrees Celsius, which significantly deteriorates its health performance. On the other hand, the main factors contributing to fuel cell system health degradation are load changing cycle, and high-power load [11]. Fig. 8 **(d)** shows the fuel cell system output power curves of the proposed strategy and Baseline B. Compared with the Baseline B not considering fuel cell SOH, the fuel cell system has significantly fewer operating points in high load aeras. The output power variance of the proposed strategy and Baseline B are 435.92 and 480.01 respectively, where the proposed strategy is 9.19% lower than that of Baseline B. The smoother load changings and fewer high-power load conditions of the proposed strategy contribute to better health performance of fuel cell system.

Quantitative data analysis can better illustrate the importance of health management. Excellent health performance is often corresponding to more efficient and reasonable operation points, which also lead to less fuel consumption. As listed in Table 5, due to the overall consideration of SOH of both power battery and fuel cells, the proposed method has the least equivalent hydrogen consumption, 6163.2 g/100 km, which decreases 2.92% compared to the Baseline B. More significantly, there is an obvious decrease of the driving cost of the proposed method. With the least SOH degradation of both power battery and fuel cells, the driving cost of the proposed method is 748.73 CNY per 100 km, which is decreased by 15.84% compared to the Baseline C.

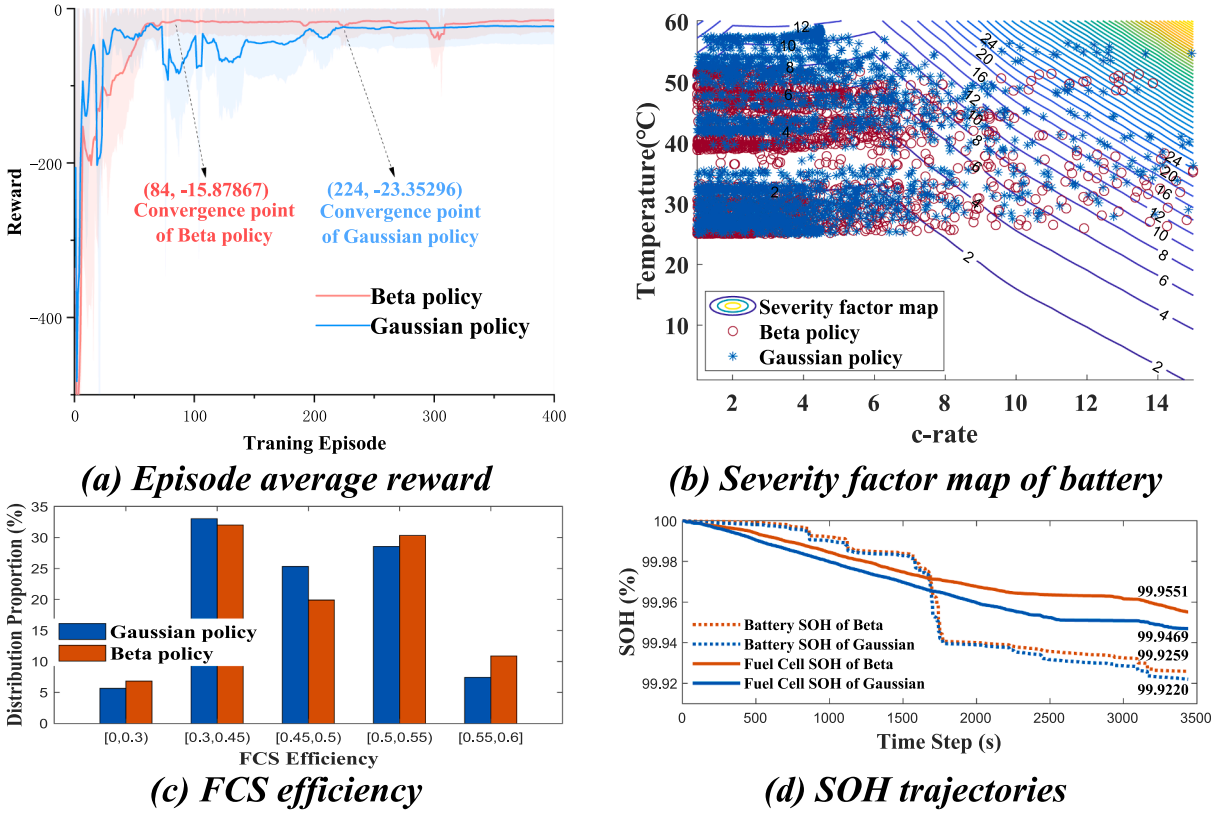

Fig. 7. Performances comparison of different stochastic policies.

Table 4
Comparison of different baselines.

Method	Battery SOH	Fuel Cell SOH
Proposed	√	√
Baseline A	×	√
Baseline B	√	×
Baseline C	×	×

In summary, the SOH constraints in reward function have proved the validity for improving health condition of both power battery and fuel cells. And they are of great significance and effectiveness for energy management system to reduce vehicle driving cost.

4.4. Analysis of optimality

In order to show the optimality of the proposed strategy, horizontal comparation experiments with other three algorithms are conducted in this section. Note that the hyper parameters of DRL-based EMSs are totally same. Fig. 9**(a)** presents SOH end values of different strategies. As we can see, the health performance of fuel cells of DP-based EMS is much better than others, but the SOH of battery is a little bad. This is because fuel cells are much more expensive than power battery, while the DP-EMS finds the global optimal solution by weighing advantages and disadvantages in the process of backward calculation. As a discrete control method, the DQN performs naturally worse than DDPG and SAC. While the health performance of DDPG-EMS is just a litter worse than that of the SAC-EMS. One of the important tasks of EMS is to maintain a reasonable and stable SOC trajectory. Taking the DP-EMS as benchmark, as shown in Fig. 9**(b)**, the DDPG-EMS and SAC-EMS can both maintain very good SOC margin during cycle. But the SOC of DQN-EMS oscillates very violently, even is very close to the upper and lower limits successively during its travel. Table 6 presents a quantitative comparison of various methods. Interestingly, while the DP-EMS achieves the best fuel cell health performance, it does not achieve the lowest equivalent hydrogen consumption due to relative neglect of power battery health. The equivalent hydrogen consumption of SAC-EMS is only 6163.2 g/100 km, which is 4.72% less than that of the DP-EMS. The equivalent hydrogen consumption of DDPG-EMS is much higher, possibly due to inappropriate operating points. In terms of money cost, the DP-EMS spends the minimum money, and the SAC-EMS is the second least which is 94.88% of DP-EMS. In summary, the proposed SAC-EMS outperforms DQN-EMS and DDPG-EMS, which are two classical DRL methods. And the proposed SAC-EMS has a 5.12% performance gap with DP-EMS in terms of money cost during driving, but is 4.72% better than DP-EMS regarding to equivalent hydrogen consumption.

4.5. Analysis of adaptability

While the massive hyper parameters of networks can be trained and updated on the cloud servers, the learned strategy still need to be executed in vehicle terminal. Vehicles will face ever changing situations during driving, so the adaptability of the proposed strategy is extremely important. In this section, the learned parameters are downloaded to initialize whole new networks with the same structures, and fixed during execution. The performance under the pre-constructed validation driving cycle is shown in Table 7. The proposed strategy shows very similar performance in terms of equivalent hydrogen consumption when facing unknown driving cycles. Health performance is also pretty good, especially the power battery SOH increases by 0.04%. With respect to the money cost during driving, the validation cycle is 811.26 CNY/100 km, which 8.35% higher than that of the training cycle. The above results demonstrate good adaptability of the proposed strategy.

5. Conclusion

In this paper, an energy management strategy which takes hydrogen

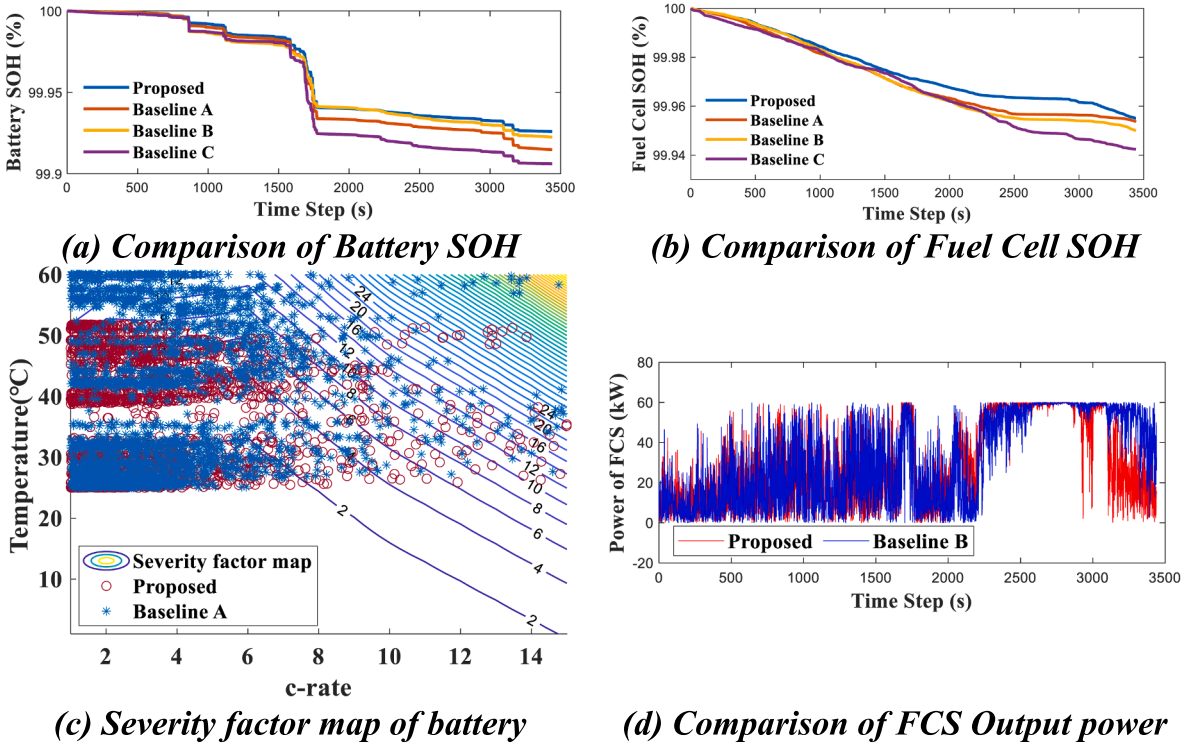

Fig. 8. Validation of health constraint.

Table 5
Performances of different methods.

Method	Equivalent H_2 cost (g/100 km)	Battery SOH (%)	Fuel Cell SOH (%)	Driving cost (CNY/100 km)	Comparison of money
Proposed	6163.2	99.9259	99.9551	748.73	84.16%
Baseline A	6253.4	99.9148	99.9535	775.75	87.18%
Baseline B	6348.8	99.9225	99.9500	825.57	92.80%
Baseline C	6298.5	99.9061	99.9424	889.64	100%

consumption, health degradation of both fuel cells and power battery, and charge margin into consideration, is proposed for FCHEV based on the improved SAC algorithm. The Beta policy with finite support is utilized to take place of the Gaussian policy with infinite support which is inconsistent with physics constraints. Main conclusions are as follows.

(1) An appropriate value of the weight coefficient is determined after extensive experiments, when $\omega = 100$, the proposed EMS realizes excellent balance among the multiple objectives.
(2) The Beta policy can choose more reasonable actions than the Gaussian policy does, and thus obtains better convergence and optimization performance.
(3) The health constraints proposed in this paper is proved to be valid for EMS of FCHEV, and it can reduce up to 15.84% in terms of money cost during driving.
(4) The proposed SAC-EMS outperforms than DQN-EMS and DDPG-EMS, and has a 5.12% performance gap with DP-EMS with

Table 6
Comparison of different EMSs.

Method	Equivalent H_2 cost (g/100 km)	Comparison of H_2	Driving cost (CNY/100 km)	Comparison of money
DP	6454.1	100%	710.37	100%
DQN	6504.1	99.23%	1137.52	62.45%
DDPG	6663.4	96.86%	777.13	91.41%
SAC	6163.2	104.72%	748.73	94.88%

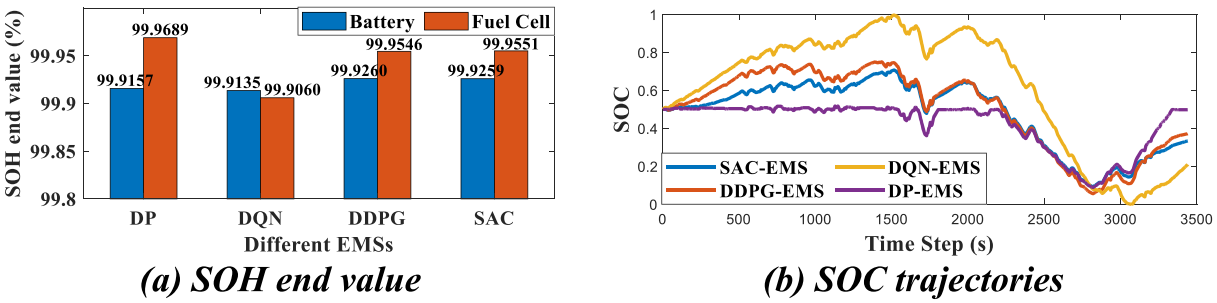

Fig. 9. Performances comparison of different EMSs.

Table 7
Performances of the proposed strategy under validation cycle.

Cycle	Equivalent H_2 cost (g/100 km)	Battery SOH (%)	Fuel Cell SOH (%)	Driving cost (CNY/100 km)	Comparison of money
Mix-train	6163.2	99.9259	99.9551	748.73	100%
Mix-valid	6102.3	99.9665	99.9677	811.26	108.35%

respect to money cost, but is 4.72% better regarding to equivalent hydrogen consumption.

(5) With respect to money cost, the validation cycle is 811.26 CNY/100 km, which 8.35% higher than that of the training cycle, demonstrating good adaptability of the proposed SAC-EMS.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (2022YFB4300300), the National Natural Science Foundation of China (Grant No.52072074), the Fundamental Research Funds for the Central Universities (Grant No.2242021R40007), Emission Peak & Carbon Neutrality Innovation S&T Project of Nanjing (No.202211018), and "Zhishan" Scholars Programs of Southeast University.

References

[1] Mohammed AS, Atnaw SM, Salau AO, et al. Review of optimal sizing and power management strategies for fuel cell/battery/super capacitor hybrid electric vehicles[J]. Energy Rep 2023;9:2213–28.
[2] He H, Wang X, Chen J, et al. Regenerative fuel cell-battery-supercapacitor hybrid power system modeling and improved rule-based energy management for vehicle application[J]. J Energy Eng 2020;146(6):04020060.
[3] Balali Y, Stegen S. Review of energy storage systems for vehicles based on technology, environmental impacts, and costs[J]. Renew Sustain Energy Rev 2021;135:110185.
[4] Fathabadi H. Novel fuel cell/battery/supercapacitor hybrid power source for fuel cell hybrid electric vehicles[J]. Energy 2018;143:467–77.
[5] Teng T, Zhang X, Dong H, et al. A comprehensive review of energy management optimization strategies for fuel cell passenger vehicle[J]. Int J Hydrogen Energy 2020;45(39):20293–303.
[6] Peng H, Li J, Thul A, et al. A scalable, causal, adaptive rule-based energy management for fuel cell hybrid railway vehicles learned from results of dynamic programming[J]. ETransportation 2020;4:100057.
[7] Wang Y, Wang L, Li M, et al. A review of key issues for control and management in battery and ultra-capacitor hybrid energy storage systems[J]. ETransportation 2020;4:100064.
[8] Sun Y, Xia C, Han J. Research on Energy Management of Fuel-Cell Electric Tractor Based on Quadratic Utility Function[J]. J Energy Eng 2023;149(1):04022044.
[9] Tang X, Jia T, Hu X, et al. Naturalistic data-driven predictive energy management for plug-in hybrid electric vehicles[J]. IEEE Trans Transp Electrif 2020;7(2):497–508.
[10] Kandidayeni M, Trovão JP, Soleymani M, et al. Towards health-aware energy management strategies in fuel cell hybrid electric vehicles: A review[J]. Int J Hydrogen Energy 2022.
[11] Tang X, Zhou H, Wang F, et al. Longevity-conscious energy management strategy of fuel cell hybrid electric Vehicle Based on deep reinforcement learning[J]. Energy 2022;238:121593.
[12] Wu G, Lee KY, Sun L, et al. Coordinated fuzzy logic control strategy for hybrid PV array with fuel-cell and ultra-capacitor in a Microgrid[J]. IFAC-PapersOnLine 2017;50(1):5554–9.
[13] Liu Y, Liu J, Zhang Y, et al. Rule learning based energy management strategy of fuel cell hybrid vehicles considering multi-objective optimization[J]. Energy 2020;207:118212.
[14] Hu X, Han J, Tang X, et al. Powertrain design and control in electrified vehicles: A critical review[J]. IEEE Trans Transp Electrif 2021;7(3):1990–2009.
[15] Ali AM, Ghanbar A, Söffker D. Optimal control of multi-source electric vehicles in real time using advisory dynamic programming[J]. IEEE Trans Veh Technol 2019;68(11):10394–405.
[16] Liu H, Xing X, Shang W, et al. NSGA-II Optimized Multiobjective Predictive Energy Management for Fuel Cell/Battery/Supercapacitor Hybrid Construction Vehicles [J]. Int J Electrochem Sci 2021;16:21046.
[17] Zhu J, Chen L, Wang X, et al. Bi-level optimal sizing and energy management of hybrid electric propulsion systems[J]. Appl Energy 2020;260:114134.
[18] Djerioui A, Houari A, Zeghlache S, et al. Energy management strategy of supercapacitor/fuel cell energy storage devices for vehicle applications[J]. Int J Hydrogen Energy 2019;44(41):23416–28.
[19] Zheng C, Cha SW, Park Y, et al. PMP-based power management strategy of fuel cell hybrid vehicles considering multi-objective optimization[J]. Int J Precis Eng Manuf 2013;14:845–53.
[20] Jiang H, Xu L, Li J, et al. Energy management and component sizing for a fuel cell/battery/supercapacitor hybrid powertrain based on two-dimensional optimization algorithms[J]. Energy 2019;177:386–96.
[21] Chen J, He H, Quan S, et al. Adaptive energy management for fuel cell hybrid power system with power slope constraint and variable horizon speed prediction [J]. Int J Hydrogen Energy 2023.
[22] Zhang W, Li J, Xu L, et al. Optimization for a fuel cell/battery/capacity tram with equivalent consumption minimization strategy[J]. Energ Conver Manage 2017;134:59–69.
[23] Li H, Ravey A, N'Diaye A, et al. Online adaptive equivalent consumption minimization strategy for fuel cell hybrid electric vehicle considering power sources degradation[J]. Energ Conver Manage 2019;192:133–49.
[24] Jinquan G, Hongwen H, Jianwei L, et al. Real-time energy management of fuel cell hybrid electric buses: Fuel cell engines friendly intersection speed planning[J]. Energy 2021;226:120440.
[25] Peng J, Chen W, Fan Y, et al. Ecological Driving Framework of Hybrid Electric Vehicle Based on Heterogeneous Multi Agent Deep Reinforcement Learning[J]. IEEE Trans Transp Electrif 2023.
[26] Xiong R, Cao J, Yu Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle[J]. Appl Energy 2018;211:538–48.
[27] Liu C, Murphey YL. Optimal power management based on Q-learning and neuro-dynamic programming for plug-in hybrid electric vehicles[J]. IEEE Trans Neural Networks Learn Syst 2019;31(6):1942–54.
[28] Liu T, Hu X, Hu W, et al. A heuristic planning reinforcement learning-based energy management for power-split plug-in hybrid electric vehicles[J]. IEEE Trans Ind Inf 2019;15(12):6436–45.
[29] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning[C]//Proceedings of the AAAI conference on artificial intelligence 2016;30(1).
[30] Han X, He H, Wu J, et al. Energy management based on reinforcement learning with double deep Q-learning for a hybrid electric tracked vehicle[J]. Appl Energy 2019;254:113708.
[31] Peng J, Fan Y, Yin G, et al. Collaborative Optimization of Energy Management Strategy and Adaptive Cruise Control Based on Deep Reinforcement Learning[J]. IEEE Trans Transp Electrif 2022.
[32] Wu Y, Tan H, Peng J, et al. Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus[J]. Appl Energy 2019;247:454–66.
[33] Li Y, He H, Peng J, et al. Deep reinforcement learning-based energy management for a series hybrid electric vehicle enabled by history cumulative trip information [J]. IEEE Trans Veh Technol 2019;68(8):7416–30.
[34] Wu C, Ruan J, Cui H, et al. The application of machine learning based energy management strategy in multi-mode plug-in hybrid electric vehicle, part I: Twin Delayed Deep Deterministic Policy Gradient algorithm design for hybrid mode[J]. Energy 2023;262:125084.
[35] Wu J, Wei Z, Li W, et al. Battery thermal-and health-constrained energy management for hybrid electric bus based on soft actor-critic DRL algorithm[J]. IEEE Trans Ind Inf 2020;17(6):3751–61.
[36] Chen W, Zhou J, Wang C, et al. Health-Aware Energy Management Strategy for Fuel Cell Hybrid Electric Vehicle Based on Soft Actor-Critic Algorithm[J]. Energy 2004;2965.
[37] Zhang H, Peng J, Tan H, et al. A deep reinforcement learning-based energy management framework with lagrangian relaxation for plug-in hybrid electric vehicle[J]. IEEE Trans Transp Electrif 2020;7(3):1146–60.
[38] Haarnoja T, Zhou A, Hartikainen K, et al. Soft actor-critic algorithms and applications[J]. arXiv preprint arXiv:1812.05905, 2018.
[39] Chou PW, Maturana D, Scherer S. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution [C]//International conference on machine learning. PMLR 2017:834–43.
[40] Pattanaik A, Tang Z, Liu S, et al. Robust deep reinforcement learning with adversarial attacks[J]. arXiv preprint arXiv:1712.03632, 2017.
[41] Lin WS, Zheng CH. Energy management of a fuel cell/ultracapacitor hybrid power system using an adaptive optimal-control method[J]. J Power Sources 2011;196(6):3280–9.

[42] Song K, Wang X, Li F, et al. Pontryagin's minimum principle-based real-time energy management strategy for fuel cell hybrid electric vehicle considering both fuel economy and power source durability[J]. Energy 2020;205:118064.
[43] Song K, Chen H, Wen P, et al. A comprehensive evaluation framework to evaluate energy management strategies of fuel cell electric vehicles[J]. Electrochim Acta 2018;292:960–73.
[44] Wei Z, Zhao D, He H, et al. A noise-tolerant model parameterization method for lithium-ion battery management system[J]. Appl Energy 2020;268:114932.
[45] Ebbesen S, Elbert P, Guzzella L. Battery state-of-health perceptive energy management for hybrid electric vehicles[J]. IEEE Trans Veh Technol 2012;61(7):2893–900.
[46] Wu J, Wei Z, Liu K, et al. Battery-involved energy management for hybrid electric bus based on expert-assistance deep deterministic policy gradient algorithm[J]. IEEE Trans Veh Technol 2020;69(11):12786–96.
[47] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International conference on machine learning. PMLR 2018:1861–70.
[48] Heess N, Wayne G, Silver D, et al. Learning continuous control policies by stochastic value gradients[J]. Adv Neural Inf Proces Syst 2015;28.
[49] Smith LN. Cyclical learning rates for training neural networks[C]//2017 IEEE winter conference on applications of computer vision (WACV). IEEE 2017:464–72.