Fast Charging Management of Lithium-Ion Battery and Cooling System: A Stackelberg Game-Based Soft Actor Critic- Deep Reinforcement Learning Method

Hongrong Yang, *Student Member, IEEE*, Quanyi Chen, *Student Member*

***Abstract*—This article proposes a fast charging management strategy for lithium-ion battery (LIB) and its cooling system, aiming to accomplish fast charging under multi-physical constraints while reducing cooling energy consumption and battery aging rates. Given the complex coupling of the LIB and the cooling system, the difficulty lies in solving the competing bilateral optimization objectives under numerous constraints. To address this challenge, we first model the problem as a Stackelberg game-based bi-level optimization, which is consistent with the fact that the charging and cooling processes are characterized by spontaneously sequential games. Then we design a corresponding Stackelberg game-based soft actor critic (SGSAC) deep reinforcement learning (DRL) method to solve the real-time gaming problem. Moreover, we prove the convergence of the proposed algorithm to ensure reliable applications. The proposed SGSAC strategy is compared experimentally with the rule-based charging strategy with proportional-integral-derivative cooling control and the state-of-art single agent DRL strategy to validate its superiority in achieving the safe and fast charging.**

***Index Terms*— Fast charging, thermal management, deep reinforcement learning (DRL), lithium-ion battery (LIB), game theory.**

I. Introduction

LITHIUM-ION batteries (LIBs) have been widely installed in electric vehicles, portable equipment and energy-storage systems, owning to the high power density and energy capacity. Fast charging technology, as one of the core technologies of LIBs, has attracted widespread attention due to its ability to alleviate range anxiety and save long charging time. However, simply increasing the charging current can cause the battery capacity degradation with accelerated aging, and even lead to severe safety issues caused by the rapidly accumulating thermal effect.

To address the dilemma, many researchers have proposed different approaches to develop new fast charging strategies, which can be divided into two types: 1) heuristic rule-based charging strategies; 2) model-based charging strategies. The well-recognized heuristic methods include the constant-current-constant-voltage (CCCV) charging or constant-power-constant-voltage (CPCV) [1], multistage constant current (MCC) [2], multistage CCCV [3], the pulse charging (PC) [4], etc. The heuristic rule-based charging protocol methods are technically mature and reliable, however, these methods highly depend on the experimental observation which is time-consuming and laborious [5]. Besides, the heuristic rule-based charging strategies are generally implicit due to the lack of consideration of internal physical and chemical processes and constraints [6].

These shortcomings of heuristic rule-based methods stimulate the emergence of the model-supported approaches which optimize the charging strategy based on different kinds of battery models [7]. They are categorized into equivalent circuit model (ECM) and electrochemical model (EM). EM has higher accuracy than ECM but requires heavier computation, which is more suitable for accurate control with pre-defined operating conditions. Conversely, ECM exhibits higher computational efficiency, which is suitable for real-time control requiring intensive training. Many researchers have made substantial effort on designing efficient model-supported charging strategies for LIBs. In [8], a constant-temperature constant-voltage closed-loop ECM-based charging technique is proposed to shorten the charging time considering the impact of temperature on the LIB capacity decay. In [9], an ECM-based LIB fast charging formation is developed to predict the electrode voltages and address the complex constrained charging optimization. In [10], an EM-based feedback control approach is proposed to achieve the faster and healthier charging without premature aging. In [11], a health-aware optimal charging technique based on EM model is designed to complete multi-physics objectives charging. Such methods are offline optimization, which use the charging trajectory generated by simulations before practical application. However, the protocols mentioned above are evaluated only for limited pre-defined operating conditions. Thus, they cannot cope with the uncertainty of the constraints in the charging process and show poor performances due to the estimation errors.

To realize the real-time accurate control for LIB fast charging, online model-based optimization methods have emerged in recent years, including proportional-integral-derivative control [12], numerical approximation solution method approach [13] - [14], model predictive control [15] - [17], etc. However, these methods have to face the following challenges:

1)The multi-objective fast charging problem is difficult to solve due to lots of parameters and intricated constraints of LIBs. Actually, most works significantly simplify the model constraints to execute online solution, leading to an inaccurate strategy.

2)The model parameters drift with battery aging, while the online model-based strategies mentioned above are not adaptive to parameter variations.

3)With the increase in the complexity of fast LIB charging problem, such as battery number, pack arrangement, cooling mode, the computing cost for real-time control will be

unbearable.

Deep reinforcement learning (DRL) is a data-driven method and can overcome these drawbacks. DRL aims at finding the optimal policy in a Markov decision process by interacting with the environment, which is naturally suited to solving real-time optimal control problems. Some researchers have made initial attempts to devise the DRL-based LIB charging strategies. Wei *et al.* [18] propose the DRL-based fast charging strategy for LIB to solve the optimal charging problem with multi-physics constraints for the first time. Park *et al.* [19] develop an DRL-based optimal-charging procedure for LIB, which has the same algorithmic architecture with that of [18]. Yang *et al.* [20] propose a soft actor critic Lagrange DRL algorithm to meet the safety constraint for LIB fast charging. The latest study is an adaptive model-based DRL approach proposed by Hao *et al.* [21], which uses Gaussian process to describe the degradation of batteries.

Nevertheless, the above works neglect the impact of cooling system which is crucial for improving the charging speed and reducing temperature creep and aging rates. In fact, the fast charging and heat management strategies are coupled deeply with gaming in a cooperative relationship. An integral fast charging management strategy should be able to perform a rapid charging safely while minimizing energy consumption and battery degradation. A few DRL-based methods have been applied in the field in charging management for electric vehicle energy system control [22] – [24], but none of the existing studies consider LIB heat dissipation, which makes the methods impractical.

This article fills the aforementioned gaps and propose a novel Stackelberg game-based fast charging management strategy for LIBs and air cooling system. The main contributions of this paper are summarized as follows:

1)This paper for the first time studies the cooperation fast charging strategy for LIB and cooling system with multi-physical consciousness. Responding to the feature of spontaneously sequential decision-making, the fast charging management problem is modeled as a bi-level optimization with Stackelberg game paradigm and then formulated as a Markov Decision Process (MDP) process.

2)A costumed Stackelberg game-based soft actor critic (SGSAC) DRL method is designed to solve the optimization problem. Moreover, we provide a theoretical proof for the convergence of the algorithm to ensure its practicability and reliability.

3)Unlike most of fast charging studies using simulation for validation, the proposed strategy is comparatively validated in real-world experiments. The results show that the proposed SGSAC strategy outperforms CCCV-PID and SAC DRL methods in terms of battery degradation, cooling energy consumption and thermal safety.

The rest of this paper is organized as follows. The mathematical formulation of fast charging management for battery and air cooling system is set forth in Section II. The SGSAC algorithm is proposed in Section III. Results and discussions are presented in Section IV and conclusion is drawn in Section V.

III. Mathematical Model

In this section, we build the mathematical model for LIB fast charging management problem comprised by battery and air-cooling system. The problem is formulated as a bi-level optimization form with sequential decisions characteristics, which is naturally the same as the Stackelberg game.

A. Lithium-Ion Battery Model

1) Electro-Thermal Model

The cylindrical lithium-iron-phosphate battery can be formulated as a coupled electro-thermal model shown in Fig 1. We use the second-order ECM to describe the battery internal dynamics, and apply a two-state thermal model to capture the battery temperature dynamics.

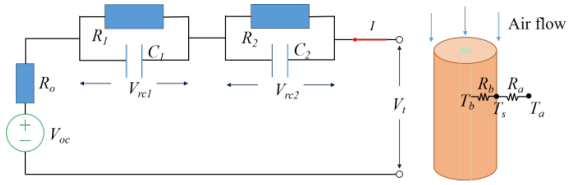

Fig. 1. Electro-Thermal model of lithium-ion battery.

Second-order ECM comprises an open circuit voltage (V_{oc}), an ohmic resistance (R_o) and two resistance-capacitor (RC) pairs (R_1, R_2, C_1, C_2). The governing equations of ECM are formulated as follows:

$$\frac{dSOC(t)}{dt} = \frac{I(t)}{Q_{batt}} \tag{1}$$

$$\frac{dV_{rc1}(SOC,T_b,t)}{dt} = -\frac{V_{rc1}(SOC,T_b,t)}{R_1(SOC,T_b,t)C_1(SOC,T_b,t)} + \frac{I(t)}{C_1(SOC,T_b,t)} \tag{2}$$

$$\frac{dV_{rc2}(SOC,T_b,t)}{dt} = \frac{I(t)}{C_2(SOC,T_b,t)} - \frac{V_{rc1}(SOC,T_b,t)}{R_2(SOC,T_b,t)C_2(SOC,T_b,t)} \tag{3}$$

$$\begin{aligned} V_t(SOC,T_b,t) = {} & V_{oc}(SOC,t) + V_{rc1}(SOC,T_b,t) \\ & + V_{rc2}(SOC,T_b,t) + I(t)R_o(SOC,T_b,t) \end{aligned} \tag{4}$$

where SOC represents the state of charge, V_{rc1} and V_{rc2} are the voltages across the RC pairs, T_b is the battery temperature, V_t is the terminal voltage, Q_{batt} is the battery nominal capacity and $I(t)$ is the applied current.

The battery temperature T_b is calculated by a two-state thermal model described as follows:

$$\frac{dT_b(t)}{dt} = \frac{T_s(t) - T_b(t)}{R_b C_b} + \frac{H_b(SOC,T_b,t)}{C_b} \tag{5}$$

$$\frac{dT_s(t)}{dt} = \frac{T_a(t) - T_s(t)}{R_a C_s} - \frac{T_s(t) - T_b(t)}{R_b C_s} \tag{6}$$

$$\begin{aligned} H_b(SOC,T_b,t) = {} & I(t)\left|V_t(SOC,T_b,t) - V_{oc}(SOC,t)\right| \\ & + I(t)(T_b + 273)E_n(SOC,t) \end{aligned} \tag{7}$$

where T_s/T_a is the surface temperature/ambient air temperature, R_b/R_a is the heat conduction resistance/convection resistance, C_b/C_s is core heat capacity/ surface heat capacity, H_b is the battery heat generation rate and E_n is the entropy change during electrochemical reactions.

2) Aging Model

Battery temperature influences the battery capacity

degradation and then affect the battery state of health (*SOH*). We adopt the aging model in [25] to calculate the SOH in charging process.

Firstly, the capacity loss in percentage Q_{loss} can be formulated as:

$$Q_{loss}(c,T_b) = H(c)A(c)^k \exp(-\frac{E_a(c)}{RT_b(t)}) \tag{8}$$

$$E_a(c) = 31700 - 370.3c \tag{9}$$

where c is the charging current rate, R is the ideal gas constant, A is the accumulated ampere-hour throughput, and k is the power-law factor. H is the pre-exponential factor as a function of C-rate, which can be found in [25]. E_a is the active energy calculated as (9).

A capacity loss of 20% is indicative of the end-of-life (EOL) of a battery. So substitute Q_{loss} = 20 into (8), we can get:

$$A(c,T_b) = [20/(H(c)\exp(-\frac{E_a(c)}{RT_b(t)}))]^{1/k} \tag{10}$$

Then the cycling number N before EOL can be calculated as:

$$N(c,T_b) = 3600A(c,T_b)/Q_{batt}. \tag{11}$$

Finally, the change of SOH is expressed as follows:

$$\frac{dSOH(c,T_b,t)}{dt} = -\frac{|I(t)|}{2N(c,T_b)Q_{batt}}. \tag{12}$$

B. Air Cooling System Model

The power of the air cooling system can be calculated as:

$$P(t) = \Delta p v_{in}(t)\rho_a A_f \tag{13}$$

where Δp is the pressure drop of the fan, v_{in} is the air velocity, ρ_a is the density of air and A_f is the cross-section area of the fan. Then we can gain the energy consumption:

$$E(t) = \int_0^{t_{ch}} P(t)\,dt \tag{14}$$

Cooling system enhances the strength heat transfer process, which is reflected by coefficient h [26]:

$$h(t) = \frac{\kappa}{D_b} Y(t)\,\mathrm{Re}(t)^{y(t)}\,\mathrm{Pr}^{1/3} \tag{15}$$

$$\mathrm{Re}(t) = \frac{\rho_a D_b v_{in}(t)}{\mu_a} \tag{16}$$

where $\kappa/Pr/\mu_a$ is the thermal conductivity/Prandtl Number/ /viscosity of the air, D_b is the diameter of a battery cell and the characteristic length for Reynolds number Re, Y and y are coefficients using look-up table.

The battery heat dissipation is influenced by the convection resistance R_a in (6), which is calculated with the convection coefficient h and the heat dissipation area S_b as follows:

$$R_a = \frac{1}{h(t)S_b} \tag{17}$$

C. Fast Charging Management

LIB fast charging management is naturally expressed as a Stackelberg game-based sequential decision-making process, where the battery acts the leader while the cooling system performs as the follower. The bi-level optimization formulation for the fast charging management problem is as follows:

$$\begin{aligned}
&\text{Leader:} && \min_{I_t} \ \nu_1 t_{ch} + \nu_2 SOH(t) \\
&&& s.t.\ I_{\min} \le I(t) \le I_{\max} \\
&&& T_{\min} \le T_b(t) \le T_{\max} \\
&&& V_{\min} \le V(t) \le V_{\max} \\
&\text{Follower:} && \min_{P_t} \ \nu_3 t_{ch} + \nu_4 E(t) \\
&&& s.t.\ P_{\min} \le P(t) \le P_{\max} \\
&&& T_{\min} \le T_b(t) \le T_{\max}
\end{aligned} \tag{18}$$

where $\nu_1 - \nu_4$ are the weights coefficients, $I_{\max}/I_{\min}$ is the upper and lower current boundary, $V_{\max}/V_{\min}$ is the maximum/minimum terminal voltage, $T_{\max}/T_{\min}$ is the maximum/minimum temperature, $P_{\max}/P_{\min}$ is the maximum/minimum power of the fan.

Although (18) is a bi-level optimization problem, it is not suitable to be solved by traditional optimization methods for the following reasons. Firstly, the problem contains a large number of nonlinear constraints, especially the calculation of SOC and terminal voltage, which makes it difficult to solve the algorithm by convex relaxation. Secondly, common optimization methods require the explicit physical battery model with accurate parameters, which is not practical in reality. Moreover, optimization methods usually have high demand on computing capacity for real-time decision making process, which imposes great burden on the controller.

III. Methodology

DRL is a data-driven method and is naturally suited to solve the dynamic, real-time control problem. To address the issues listed in section II, we propose a **S**tackelberg **g**ame based **s**oft **a**ctor **c**ritic method to solve the fast charging management problem which is formulated as a Markov decision process. Besides, we prove the convergence of the proposed method to ensure the practicable applications.

A. Stackelberg Game-Based Soft Actor Critic Algorithm

The MDP tuple in multi-agent RL can be defined as a tuple (S, A_i, T, r_i, γ), where S is the state space and $T: S \times A \to D(S)$ is the transaction distribution, A_i/r_i denotes the action space/reward function of agent i and $\gamma \in (0,1)$ is the discounter factor for measuring the long-term reward. A policy $\pi: S \to \delta(A)$ is a mapping from state to distribution over actions. In this study, we have two types of agents based on the sac algorithm, leader and follower, corresponding to the battery and the heat dissipation system, respectively. The two SAC-based agents aim at maximizing the maximum entropy objective which generalizes the standard objective extending it with an adaptive entropy term. Assuming agent 1 is the leader and agent 2 is the follower, corresponding to the bi-level optimization (18), the framework of SGSAC algorithm can be formulated as follows:

$$\begin{aligned}
&\max_{\pi_1} \ \mathbb{E}_{(\boldsymbol{s}_t, \vec{\boldsymbol{a}_t}) \sim (\rho_{\pi_1}, \rho_{\pi 2})} \sum_{t=0}^{T} \gamma^t [r_1(\boldsymbol{s}_t, \vec{\boldsymbol{a}_t}) + \alpha_1^T \mathcal{H}(\pi_1(\cdot \mid \boldsymbol{s}_t, \boldsymbol{a}'_{2,t})] \\
&\text{s.t.} \quad \pi_1 \in \Pi_1 \\
&\max_{\pi_2} \ \mathbb{E}_{(\boldsymbol{s}_t, \vec{\boldsymbol{a}_t}) \sim (\rho_{\pi_1}, \rho_{\pi 2})} \sum_{t=0}^{T} \gamma^t [r_2(\boldsymbol{s}_t, \vec{\boldsymbol{a}_t}) + \alpha_2^T \mathcal{H}(\pi_2(\cdot \mid \boldsymbol{s}_t, \boldsymbol{a}_{1,t})] \\
&\text{s.t.} \quad \pi_2 \in \Pi_2
\end{aligned} \tag{19}$$

where $\overrightarrow{\boldsymbol{a}_t} = (\boldsymbol{a}_{1,t}, \boldsymbol{a}_{2,t})$ is the joint vector of actions, $\boldsymbol{a}'_{2,t}$ is the is the opponent's action predicted by its strategy, which is shown as (27). ρ_π is the state-action marginal trajectory distribution under the policy π. α_i^T is the temperature hyperparameter that adjusts the weight of entropy $\mathcal{H}$ to reward.

The Bellman function that describes the iterative relationship between the soft state-values $V_i(\boldsymbol{s}_t)$ and soft action-values $Q_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t})$:

$$Q_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) = r_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) + \gamma \mathbb{E}_{\boldsymbol{s}_{t+1} \sim p}\left[V_i(\boldsymbol{s}_{t+1})\right] \tag{20}$$

where

$$V_i(\boldsymbol{s}_t) = \mathbb{E}_{\boldsymbol{a}_{i,t} \sim \pi_i}\left[Q_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) - \alpha_i^T \log \pi_i(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)\right] \quad i = 1,2 \tag{21}$$

where $\widetilde{\boldsymbol{a}_t}$ denotes the action of the other agent. Each agent simultaneously learns a policy π_{φ_i}, a Q-function Q_{θ_i} and a target Q-function $Q_{\overline{\theta}_i}$ using a policy network and four critic networks, respectively.

The critic network takes the joint action as its input in addition to the current state and the soft Q-function parameters can be learned by minimizing the Bellman residual:

$$J_Q(\theta_i) = \mathbb{E}_{(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) \sim \mathcal{D}}[\frac{1}{2}(Q_{\theta_i}(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) - (r_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) + \gamma \mathbb{E}_{\boldsymbol{s}_{t+1} \sim p}[V_{\overline{\theta}_i}(\boldsymbol{s}_{t+1})]))^2] \quad i = 1,2 \tag{22}$$

The gradient for θ_i can be optimized through stochastic gradients as:

$$\begin{aligned}\nabla_{\theta_i} J_Q(\theta_i) &= \nabla_{\theta_i} Q_{\theta_i}(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t})(Q_{\theta_i}(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) - (r_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) \\ &+ \gamma(Q_{\overline{\theta}_i}(\boldsymbol{s}_{t+1}, \overrightarrow{\boldsymbol{a}_{t+1}}) - \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t+1} \mid \boldsymbol{s}_{t+1}, \widetilde{\boldsymbol{a}_{t+1}})))\ i = 1,2\end{aligned} \tag{23}$$

where parameters $\overline{\theta}_i$ are obtained as an exponentially moving average of the soft Q-function weights.

The policy network observes both the current state and the opponent's action. The policy parameters can be learned by minimizing the Kullback-Leibler divergence:

$$J_\pi(\varphi_i) = \mathbb{E}_{(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}) \sim \mathcal{D}}\left[\mathbb{E}_{\boldsymbol{a}_{i,t} \sim \pi_i}\left[\alpha_i^T \log \pi_i(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) - Q_i(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t})\right]\right] \quad i = 1,2 \tag{24}$$

We use the reparameterization trick to optimize φ_i because our target is the Q-function:

$$\begin{aligned}\nabla_\varphi J_{\pi_i}(\varphi_i) = &\nabla_{\varphi_i} \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)) \\ &+ (\nabla_{\boldsymbol{a}_{i,t}} \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)) \\ &- \nabla_{\boldsymbol{a}_{i,t}} Q(\boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t}))\nabla_{\varphi_i} f_{\varphi_i}(\varepsilon_{i,t}; \boldsymbol{s}_t, \boldsymbol{a}_t) \quad i = 1,2\end{aligned} \tag{25}$$

where $\varepsilon_{i,t}$ is an input noise vector which is sampled from Gaussian distribution in this study.

Then we have the following update rules for Stackelberg game:

$$\boldsymbol{a}_{1,t+1} = \pi_{\varphi_1}(\cdot \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}'_{2,t+1}) \tag{26}$$

$$\boldsymbol{a}'_{2,t+1} = \pi_{\varphi_2}(\cdot \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}_{1,t}) \tag{27}$$

$$\boldsymbol{a}_{2,t+1} = \pi_{\varphi_2}(\cdot \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}_{1,t+1}) \tag{28}$$

$$\theta_i = \theta_i - \beta_Q \nabla_{\theta_i} J_Q(\theta_i) \quad i = 1,2 \tag{29}$$

$$\varphi_i = \varphi_i - \beta_{\pi_i} \nabla_\varphi J_{\pi_i}(\varphi_i) \quad i = 1,2 \tag{30}$$

$$\overline{\theta}_i = \tau\theta_i + (1-\tau)\overline{\theta}_i \quad i = 1,2 \tag{31}$$

$$\alpha_i^{T*} = \arg\min_{\alpha_i^T} \mathbb{E}_{\boldsymbol{a}_{i,t} \sim \pi_i}\left[-\alpha_i^T \log \pi_i(\boldsymbol{a}_{i,t} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) - \alpha_i^T \overline{\mathcal{H}_i}\right] \quad i = 1,2 \tag{32}$$

where $\overline{\mathcal{H}}$ is the desired minimum expected entropy, the temperature hyperparameter is updated simultaneously with critic networks. Leader understands the policy of follower, for each environment step, the leader first anticipates the follower's action confronting the new state, and then sample the action based on the follower's prediction action and new state as (26) - (27). After that, follower makes action based on the identified leader's action and state as (28). For each gradient step, the two agents iteratively update the critic parameters/ policy parameters/ target network parameters/ temperature hyperparameter according to (29)/(30)/(31)/(32). The detailed SGSAC algorithm is presented as follows:

Algorithm 1 : Proposed SGSAC Algorithm

Input: $\theta_1^{\mathrm{I}}, \theta_1^{\mathrm{II}}, \overline{\theta}_1^{\mathrm{I}}, \overline{\theta}_1^{\mathrm{II}} / \theta_2^{\mathrm{I}}, \theta_2^{\mathrm{II}}, \overline{\theta}_2^{\mathrm{I}}, \overline{\theta}_2^{\mathrm{II}}$ -initial parameters of four critic networks of leader/follower; φ_1 / φ_2 -initial parameters of the policy network of leader/follower; $\mathcal{D} \leftarrow \varnothing$ -initialize an empty replay buffer; γ -discount factor; $\beta_{Q_1}, \beta_{\varphi_1}, \beta_{\alpha_1} / \beta_{Q_2}, \beta_{\varphi_2}, \beta_{\alpha_2}$ -learning rate of critic network, policy network and temperature coefficient of leader/follower.

for each iteration **do**
  for each environment step **do**
    Leader estimates the policy of follower confronting the new state $\pi_{\varphi_2}(\cdot \mid \boldsymbol{s}_{t+1}, \boldsymbol{a}_{1,t})$ and then give the leading action based on (26)-(27).
    Follower chooses the optimal action in face of the determined status based on (28).
    Sample transition from the environment $\boldsymbol{s}_{t+1} \sim p(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \overrightarrow{\boldsymbol{a}_t})$
    Store the transition in replay buffer $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{s}_t, \boldsymbol{a}_{1,t}, \boldsymbol{a}_{2,t}, \boldsymbol{s}_{t+1}, \boldsymbol{r}_{1,t+1}, \boldsymbol{r}_{2,t+1})\}$
  end for
  for each gradient step **do**
    Update critic parameters for the two agents based on (22), (23), (29).
    Update policy weights based for the two agents based on (24), (25), (30).
    Update target network weights for the two agents based on (31).
    Adjust temperature for the two agents based on (32).
  end for
end for

Output: $\theta_1^{\mathrm{I}}, \theta_1^{\mathrm{II}}, \overline{\theta}_1^{\mathrm{I}}, \overline{\theta}_1^{\mathrm{II}} / \theta_2^{\mathrm{I}}, \theta_2^{\mathrm{II}}, \overline{\theta}_2^{\mathrm{I}}, \overline{\theta}_2^{\mathrm{II}}$, φ_1 / φ_2

B. Convergence Proof of SGSAC

Assumption 1: *Every state-action pair* $(s_t, a_{1,t}, a_{2,t})$ *is visited infinitely often.*

Assumption 2: *We define the learning rate* $\beta_t(s_t, a_{1,t}, a_{2,t})$ *as the inverse of the number of times that the state-action pair* $(s_t, a_{1,t}, a_{2,t})$ *has been visited for learning.* κ_t *satisfies the following conditions:*

1) $0 \le \beta_t(s_t, a_{1,t}, a_{2,t}) < 1$, $\sum_{t=0}^{\infty} \beta_t(s_t, a_{1,t}, a_{2,t}) = \infty$, $\sum_{t=0}^{\infty}[\beta_t(s_t, a_{1,t}, a_{2,t})]^2 < \infty$, *the latter two hold uniformly and with probability 1.*

2) $\beta_t(s, a_1, a_2) = 0$ *if* $(s, a_1, a_2) \neq (s_t, a_{1,t}, a_{2,t})$, *which means the agent updates through Q-function.*

Assumption 3: *Each stage of Stackelberg game has a global optimal point, and the rewards of this point are selected by the actor function to update the soft Q-function with probability 1.*

Assumption 3 is difficult to satisfied when the two agents are competing completely. Thus, the expected relationship between leader and follower is cooperative.

Our convergence proof relies on the lemma 1 by Szepesvari and Littman [28] as follows:

Lemma 1: *Assume that* β_t *satisfies the Assumption 2 and the mapping* $P_t: \mathbb{Q} \to \mathbb{Q}$ *satisfies the following conditions:*

1) there exists a number $0 < \gamma < 1$ *and a sequence* $\lambda_t \geq 0$ *converging to zero with probability 1 such that* $\|P_t Q - P_t Q^*\| \leq \gamma \|Q - Q^*\| + \lambda_t$ *for all* $Q \in \mathbb{Q}$ *and*

2) $Q^* = E[P_t Q^*]$,

then the iteration defined by

$$Q_{t+1} = (1-\beta_t)Q_t + \beta_t[P_t Q_t] \tag{33}$$

converges to Q^* *with probability 1.*

First we prove the convergence point satisfies the *condition 1* of *lemma 1*. We give two definitions:

Definition 1: *Let* $Q = (Q_1, Q_2)$, *where* $Q_1 \in \mathbb{Q}_1, Q_2 \in \mathbb{Q}_2$, *and* $\mathbb{Q} = \mathbb{Q}_1 \times \mathbb{Q}_2$. P_t *is a mapping on the complete metric space* $\mathbb{Q} \to \mathbb{Q}$, $P_t Q = (P_t Q_1, P_t Q_2)$, *where*

$$P^t Q_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) = r_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \gamma Q_i(\boldsymbol{s}_{t+1}, \vec{\boldsymbol{a}}_{t+1}) \quad i = 1,2 \tag{34}$$

Definition 2:

$$\begin{aligned} \left\|Q - Q\right\| &\equiv \max_j \max_{\boldsymbol{s}_t} \left\|Q^j(\boldsymbol{s}_t) - Q^j(\boldsymbol{s}_t)\right\|_{(i,\boldsymbol{s}_t)} \\ &\equiv \max_j \max_{\boldsymbol{s}_t} \max_{\vec{\boldsymbol{a}}_t} \left|Q^j(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) - Q^j(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)\right| \end{aligned} \tag{35}$$

Lemma 2 ((Hu and Wellman [29]), Lemma 16): $\|P^t Q - P^t \hat{Q}\| \leq \gamma \|Q - \hat{Q}\|, \forall Q, \hat{Q} \in \mathbb{Q}$.

According to *lemma 2*, $\|P^t Q - P^t \hat{Q}\| \leq \gamma \|Q - \hat{Q}\| \leq \gamma \|Q - Q^*\| + \lambda_t, \lambda_t \geq 0$, where P^t is a contraction operator.

Then we prove *condition 2* of *lemma 1*, when there is a convergence point, the following equation occurs:

$$\begin{aligned} &Q_i^*(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) \\ &= r_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \gamma \sum_{\boldsymbol{s}_{t+1} \in \boldsymbol{S}} p(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)(Q_i^*(\boldsymbol{s}_{t+1}, \vec{\boldsymbol{a}}_{t+1}) + \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t+1} \mid \boldsymbol{s}_{t+1}, \tilde{\boldsymbol{a}}_{t+1}))) \\ &= \sum_{\boldsymbol{s}_{t+1} \in \boldsymbol{S}} p(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)(r_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \gamma(Q_i^*(\boldsymbol{s}_{t+1}, \vec{\boldsymbol{a}}_{t+1}) + \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t+1} \mid \boldsymbol{s}_{t+1}, \tilde{\boldsymbol{a}}_{t+1})))) \\ &= \mathbb{E}_{\theta_i}[P^t Q_i^*(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)] \qquad i = 1,2 \end{aligned} \tag{36}$$

Then, we have:

$$\begin{aligned} Q^*(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) &= (Q_1^*(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t), Q_2^*(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)) \\ &= \mathbb{E}_{\theta_i}[P^t(Q_1^*(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t), Q_1^*(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t))] \\ &= \mathbb{E}[P^t Q^*(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t)] \end{aligned} \tag{37}$$

Finally, rewrite (31) to the form expressed by soft Q-function:

$$\begin{aligned} Q_i(\boldsymbol{s}_{t+1}, \vec{\boldsymbol{a}}_{t+1}) = (1-\beta_t)Q_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \beta_t(r_i(\boldsymbol{s}_t, \vec{\boldsymbol{a}}_t) + \\ \gamma(Q_i^*(\boldsymbol{s}_{t+1}, \vec{\boldsymbol{a}}_{t+1}) + \alpha_i^T \log(\pi_{\varphi_i}(\boldsymbol{a}_{i,t+1} \mid \boldsymbol{s}_{t+1}, \tilde{\boldsymbol{a}}_{t+1})))) \; i = 1,2 \end{aligned} \tag{38}$$

(38) conforms to the format of (44) and satisfies the *condition 1 and 2* in *lemma 1*, so the Q value of SGSAC will converge to Q^* with probability 1 and the actor function will be trained correctly under the converged Q function.

C. RL Formulation of the Fast Charging Problem

In this section, the fast charging management problem is formulated as a MDP conforming to RL reformulation. The framework of the SGSAC strategy for fast charging management is shown schematically in Fig. 2.

1) Definition of State

s_t is the observed status information of battery and cooling system, including the temperature $T(t)$, the terminal voltage $V(t)$, the state of charge $SOC(t)$.

2) Definition of Action

$a_{1,t}$ is the charging current $I(t)$ of battery while $a_{1,t}$ is the cooling system power $P(t)$. The values of the actions should be pre-constrained according to priori knowledge to save the computation time.

3) Definition of Reward and Cost

The reward functions of two agents are given as follows with positive weights $\omega_1 - \omega_6$:

Agent 1 (Battery):

$$J_1(t) = \omega_1 C_{fast}(t) + \omega_2 C_{soh}(t) + \omega_3 C_{volt}(t) + \omega_4 C_{temp}(t) \tag{39}$$

$$C_{fast}(t) = -\left|SOC_{tar} - SOC(t)\right| \tag{40}$$

$$C_{soh}(t) = SOH(t) - SOH(t-1) \tag{41}$$

$$C_{volt}(t) = \begin{cases} 0 & , V_{\min} \leq V(t) \leq V_{\max} \\ V_{\max} - V(t), & V(t) > V_{\max} \\ V(t) - V_{\min}, & V(t) > V_{\min} \end{cases} \tag{42}$$

$$C_{temp}(t) = \begin{cases} 0 & , T_b(t) \leq T_{\max} \\ T_{\max} - T_b(t), & T_b(t) > T_{\max} \end{cases} \tag{43}$$

Agent 2 (Air cooling system):

$$J_2(t) = \omega_5 C_{fast}(t) + \omega_6 E(t) + \omega_7 C_{temp}(t) \tag{44}$$

where $C_{fast}(t)$ is the cost for incentive fast charging, which is formulated as (40), SOC_{tar} is the target SOC. $C_{soh}(t)$ is the battery degradation cost shown as (41) and $C_{volt}(t)/C_{temp}(t)$ denotes the cost of voltage/temperature, which is shown as (42)/(43).

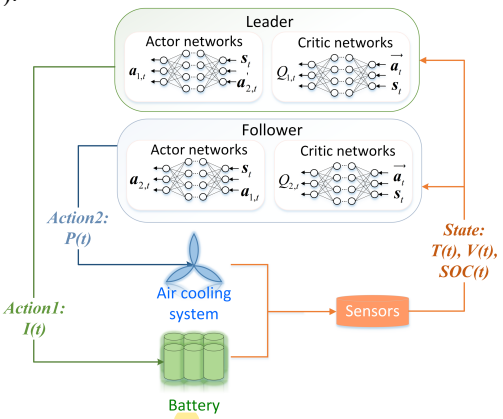

Fig. 2 Implementation of SGSAC-based fast charging management strategy.

IV Results And Discussions

A. Model Validation

The A123 26650 LIB cell is cycled with 2C, 4C, 6C by Neware testing system, which supports accurate current output and high speed data logging. The average of two thermocouples affixed to the axial midpoint of the cell is taken as the surface temperature and recorded by the data collector. We apply the modified ECM parameters in [25] as the LIB detailed parameters. During the experimental validation, the ambient temperature is 23.5 °C and the humidity is 40%. The comparative results in terms of temperature and terminal voltage are given in Fig.3. The statistics LIB modeling errors which are described by root mean square error (RMSE) under different charging rates are presented in Table I. According to the existing LIB parameters identification in [30] - [31], the LIB

modeling errors are acceptable, i.e., voltage RMSE ⩽ 50 mV and temperature RMSE ⩽ 1°C. According to the results of RMSE in Table II, the modeling errors under different charging rates are all in acceptable range.

We measure the power of brushless fan and the corresponding wind velocity at the inlet of the test device to define the air cooling system characteristics. To avoid interference of the fan's time-vary internal resistance with the energy consumption results, we collect wind speed data from the suction side. The detailed results are given in Fig. 4.

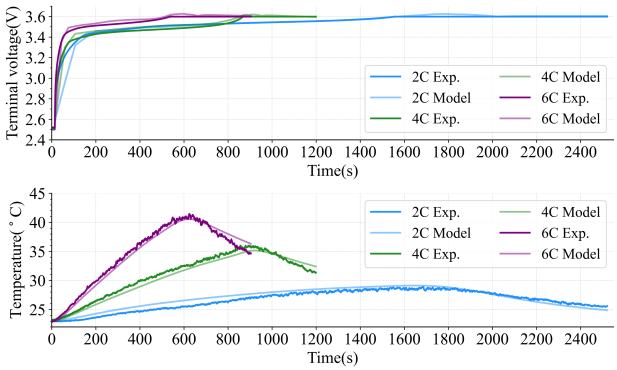

Fig. 3 Results of lithium-ion battery parameters validation in terms of voltage and temperature output.

TABLE I
MODELING ERRORS OF DIFFERENT C-RATES

Item	Terminal voltage (V)			Surface temperature (°C)		
Rate	2C	4C	6C	2C	4C	6C
RMSE	0.0464	0.0429	0.0412	0.6288	0.6767	0.7689

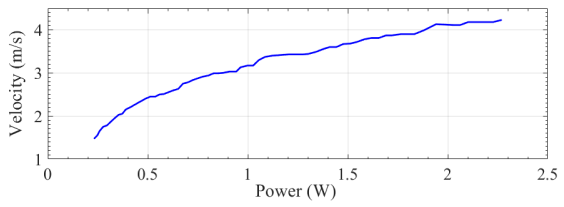

Fig. 4 Relationship between fan power and air velocity at test device inlet.

B. *Training and Simulation*

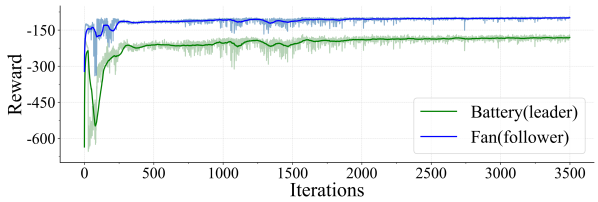

Fig. 5 Reward of training process.

The training results of the two-stage training process are shown in Fig.3, in which the darker line represents the average value per 50 steps while the lighter line denotes the actual value of each training step. The specific parameters of SGSAC are given in Table II.

TABLE II
ALGORITHM PARAMETERS

Parameters	Value
Optimizer	Adam
Number of hidden layers (All networks)	2
Number of hidden units per layer of policy network	64/512
Number of hidden units per layer of critic networks	256/16
Learning rate of actor network/critic network/temperature coefficient	5e-4/1e-3/1e-3
Discount factor	0.98
Replay buffer size	1e6
Number of samples per minibatch	256
Nonlinearity	Leaky-ReLU

To eliminate the influence of model error and conduct a theoretical validation, we first test the trained policy in a simulation environment. The comparative results are given in Fig. 6, where the charging strategy and cooling method of the control group are CCCV and proportional-integral-derivative (PID) control. The initial ambient temperature is 33°C, and the maximum safe temperature is set at 41°C according to the operating safe temperature in [32]. Considering the risk of overcharging in experiment, the target SOC for DRL is set as 97%. Fig. 6 clearly indicates that the SGSAC strategy succeeds to balance the charging speed, charging limits and energy consumption of fan. Conversely, due to the hysteresis effect and the delayed nature of the temperature drop, despite the PID target temperature is set less than 41°C, the traditional CCCV-PID methods violates the temperature constraint and waste lots of energy under 6 charging rates in simulation.

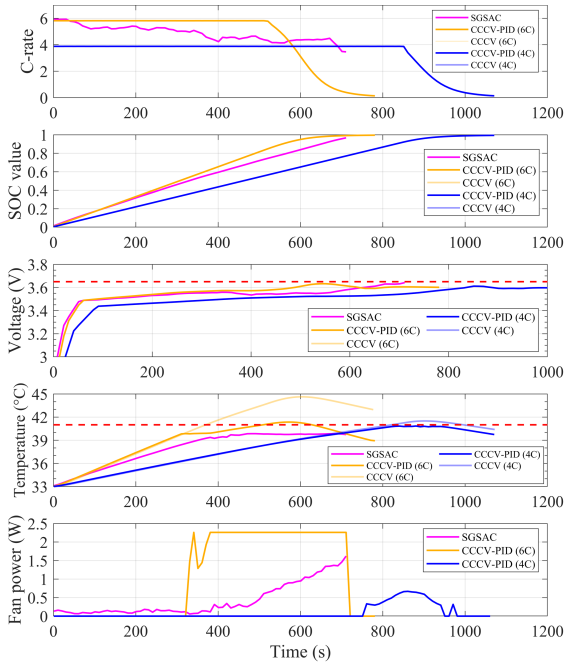

Fig. 6 Simulation results of SGSAC method with traditional strategies.

Fig. 7 shows the comparison of the SGSAC and SAC DRL method whose reward function is the sum of the two agents from SGSAC. The simulation results indicate that neither the SGSAC nor SAC strategies bleach the charging constraints, the charging speed of both is basically the same, but SGSAC has a much better performance in energy consumption. In fact, we spend a lot of time tuning the SAC algorithm, much more than the SGSAC, because of the adjustment for the balance of so many objectives in reward functions. In contrast, SGSAC is more practical due to the ease of convergence of Stackelberg game and few items of reward functions for each agent.

Compare with the SAC and CCCV-PID(6C), SGSAC has the most uniform charging current and lowest power consumption, attributed to the benign game between the two agents. The comparison results show that neither the SAC or CCCV-PID

truly considers the benefit of the cooling system, i.e., the energy cost is completely yielding to the charging speed, which makes their cooling strategies very radical, leading to the unnecessary energy waste and risk of violation. Especially, temperature violation can accelerate the battery degradation according to the aging model.

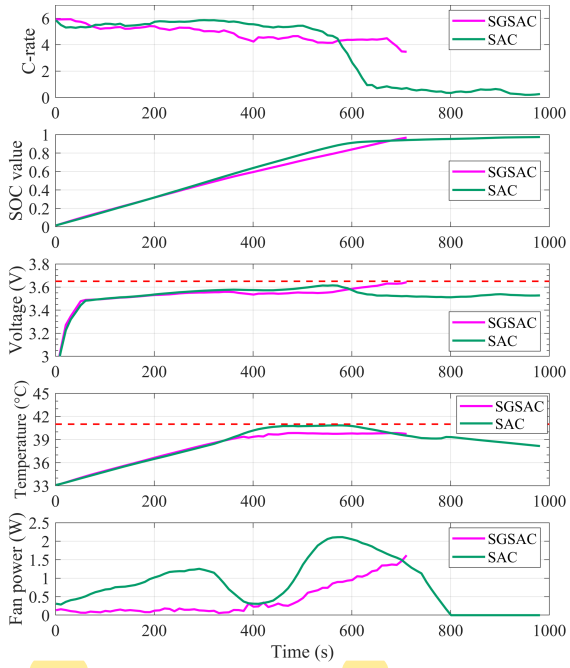

Fig. 7 Simulation results of SGSAC with SAC DRL strategy.

A simulation of 100 charging cycles is executed to evaluate the influence of the prooposed strategy on battery degradation. The results of SoH drop using different charging strategies are given in Table III. It is shown that the SGSAC strategy has the lowest battery degradation. Compared to SAC/CCCV-PID(6C)/CCCV-PID(4C), SGSAC extends LIB life by 2.8%/12.2%/0.3%. Finally, the execution time for performing different strategies are shown in Table IV. The results show that each strategy is effective enough for real-time execution.

TABLE III
SOH DROPS FOR 100 CHARGING CYCLES USING DIFFERENT STRATEGIES

Strategy	SGSAC	SAC	CCCV-PID(6C)
SOH drop	1.589%	1.633%	1.782%
Strategy	CCCV(6C)	CCCV-PID(4C)	CCCV(4C)
SOH drop	1.823%	1.593%	1.644%

TABLE IV
EXECUTION TIME PER STEP OF DIFFERENT STRATEGIES

Strategy	SGSAC	SAC	CCCV-PID(6C)
Time(ms)	9.93	6.75	3.98

C. *Experimental Validation*

In this section, the proposed strategy is applied for real-world battery fast charging management problem and further compared with different methods. The experiment is executed in the built test environment shown in Fig. 8 to evaluate the performance of the trained strategy. The test device is a custom sealed box which is bale to prevent the influence of the airflow from the thermal chamber's fan. The suction side of the fan is embedded in the case of the test device, while the battery is held in the case of the test device being concentric with the fan.

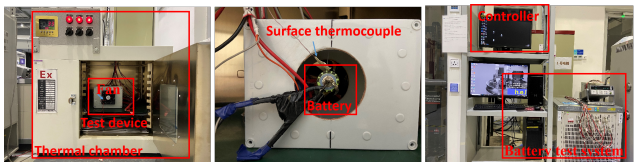

Fig. 8 Experiment setup

The experimental results by using CCCV, CCCV-PID, SGSAC strategy are comparatively shown in Fig. 9. It is obvious that the experimental results generally coincide with the conclusions of simulation. None of the strategies induces voltage violation, however, the temperature under the traditional CCCV strategy goes beyond the limitation greatly, CCCV-PID strategy is effective at 4C rate, but no longer possible at 6C rate. Besides, SGSAC dominates in terms of the thermal management reflecting by the lowest maximum and average temperatures in stabilization period. Fig. 10 presents the experimental results of SGSAC and SAC strategies. Compared to SAC strategy, SGSAC strategy charges quicker and offer a better controllability in temperature and voltage constraints with fewer energy consumption.

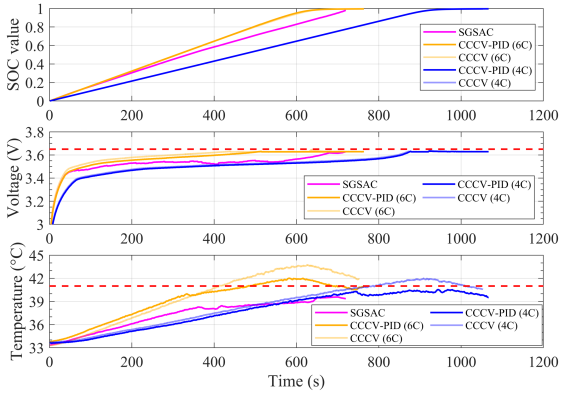

Fig. 9 Experimental results of SGSAC method with CCCV-PID strategies.

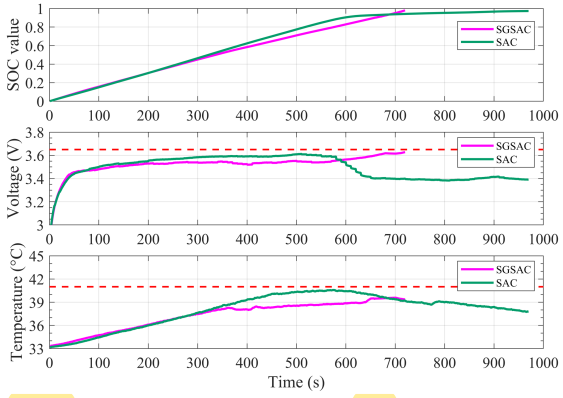

Fig. 10 Experimental results of SGSAC with SAC DRL strategy.

The detailed quantitative comparisons in terms of battery charging speed and fan energy consumption are given in Table V and VI, respectively. The SGSAC strategy presents the smart management of the trade-off between the charging speed and energy consumption under the operating constraints. Compared to the CCCV-PID(4C), SGSAC reduces 21.17%, 21.19% charging time for 90%, 97% SOC charging end points; compared to SAC strategy, SGSAC reduces 23.23% charging

time and saves 64.68% energy consumption during the full charging process.

TABLE V
COMPARISON OF CHARGING SPEED OF DIFFERENT STRATEGIES

Strategy	SGSAC	SAC	CCCV-PID(6C)	CCCV-PID(4C)
To 90% SOC	659s	595s	561s	836s
To 97% SOC	714s	930s	620s	906s

TABLE VI
COMPARISON OF FAN ENERGY CONSUMPTION OF DIFFERENT STRATEGIES

Strategy	SGSAC	SAC	CCCV-PID(6C)	CCCV-PID(4C)
To 90% SOC	215.4J	554.2J	513.3J	27.9J
To 97% SOC	291.9J	826.4J	496.0J	72.9J

V. CONCLUSION

A Stackelberg game-based charging strategy is proposed for LIB and cooling system to solve the fast charging management problem with thermal and voltage safety constraints. A novel SGSAC DRL algorithm is designed to tradeoff the bilateral benefits through the competing and reciprocal gaming. We can summarize the following conclusions:

1) SGSAC strategy charge the LIB at 97% SOC in 291.9s and consume 291.9 J energy without violating any constraints.

2) SGSAC strategy has the lowest battery degradation and extends LIB life by 2.8%/12.2%/0.3%, comparing to SAC/CCCV-PID(6C)/CCCV-PID(4C), SGSAC.

3)SGSAC exhibits the most stable temperature performance reflecting by the lowest maximum and average temperatures in the stabilization period.

4)SGSAC performs well in charging speed and energy consumption. Compared to CCCV-PID (4C), SGSAC reduces 21.19% charging time; compared to SAC, SGSAC reduces 23.23% charging time and 64.75% energy consumption.

REFERENCES

[1] K. Liu, C. Zou, K. Li, and T. Wik, "Charging pattern optimization for lithium-ion batteries with an electrothermal-aging model," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5463–5474, Dec. 2018.

[2] H. Bizhani, S. K. H. Sani, H. Rezazadeh, S. M. Muyeen and S. Rahmani, "Modeling of an optimum fast charging multi-step constant current profile for lead-acid batteries," *IEEE Trans. Ind. Appl.*, vol. 59, no. 2, pp. 2050-2060, Mar./Apr. 2023.

[3] D. Anseán et al., "Fast charging technique for high power lifepo4 batteries: A mechanistic analysis of aging," *J. Power Sources*, vol. 321, pp. 201–209, 2016.

[4] X. Huang, W. Liu, A. B. Acharya, J. Meng, R. Teodorescu, and D.-I. Stroe, "Effect of pulsed current on charging performance of lithium-ion batteries," *IEEE Trans. Ind. Electron.*, vol. 69, no. 10, pp. 10144–10153, Oct. 2022.

[5] L. R. Chen, "Design of duty-varied voltage pulse charger for improving Li-ion battery-charging response," *IEEE Trans. Ind. Electron.*, vol. 56, no. 2, pp. 480–487, Feb. 2009. C. Chen, Z. Wei, and A. C. Knoll, "Charging optimization for Li-ion battery in electric vehicles: A review," *IEEE Trans. Transp. Electrific.*, vol. 8, no. 3, pp. 3068–3089, Sep. 2022.

[7] Nikolaos Wassiliadis Jakob Schneider, Alexander Frank, LeoWildfeuer, Xue Lin, Andreas Jossen, Markus Lienkamp, "Review of fast charging strategies for lithium-ion battery systems and their applicability for battery electric vehicles", *J. Energy Storage*, vol. 44, Dec. 2021.

[8] L. Patnaik, A. V. J. S. Praneeth, and S. S. Williamson, "A closed-loop constant-temperature constant-voltage charging technique to reduce charge time of lithium-ion batteries," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1059–1067, Feb. 2019.

[9] Drees, Robin, Frank Lienesch, and Michael Kurrat. "Fast charging lithium-ion battery formation based on simulations with an electrode equivalent circuit model." *J. Energy Storage*, vol. 36, Apr. 2021.

[10] L. D. Couto et al., "Faster and healthier charging of lithium-ion batteries via constrained feedback control," *in IEEE Trans. Control Syst. Technol.*, vol. 30, no. 5, pp. 1990-2001, Sep. 2022.

[11] Y. Gao et al., "Health-aware multiobjective optimal charging strategy with coupled electrochemical-thermal-aging model for lithium-ion battery," *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 3417–3429, May 2020.

[12] Z. Chu, X. Feng, L. Lu, J. Li, X. Han, and M. Ouyang, "Non-destructive fast charging algorithm of lithium-ion batteries based on the control oriented electrochemical model," *Appl. Energy*, vol. 204, pp. 1240–1250, 2017.

[13] Nambisan P, Saha P, Khanra M. "Real-time optimal fast charging of Li-ion batteries with varying temperature and charging behaviour constrains," *J. Energy Storage*, vol. 41, Sep. 2021.

[14] D.-J. Xuan et al., "Real-time estimation of state-of-charge in lithium-ion batteries using improved central difference transform method," *J. Cleaner Prod.*, vol. 252, Apr. 2020.

[15] N. Tian, H. Fang, and Y. Wang, "Real-time optimal lithium-ion battery charging based on explicit model predictive control," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 1318–1330, Feb. 2021.

[16] J. Liu, G. Li, and H. K. Fathy, "An extended differential flatness approach for the health-conscious nonlinear model predictive control of lithium-ion batteries," *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 5, pp. 1882–1889, Sep. 2017.

[17] C. Zou, X. Hu, Z. Wei, T. Wik, and B. Egardt, "Electrochemical estimation and control for lithium-ion battery health-aware fast charging," *IEEE Trans. Ind. Electron.*, vol. 65, no. 8, pp. 6635–6645, Aug. 2018.

[18] Z. Wei, Z. Quan, J. Wu, Y. Li, J. Pou, and H. Zhong, "Deep deterministic policy Gradient-DRL enabled multiphysics-constrained fast charging of lithium-ion battery," *IEEE Trans. Ind. Electron.*, vol. 69, no. 3, pp. 2588–2598, Mar. 2022.

[19] S. Park et al., "A deep reinforcement learning framework for fast charging of Li-ion batteries," *IEEE Trans. Transport. Electrific.*, vol. 8, no. 2, pp. 2770–2784, Jun. 2022.

[20] X. Yang, *et al.* "Enabling safety-enhanced fast charging of electric vehicles via soft actor Critic-Lagrange DRL algorithm in a Cyber-Physical system," *Applied Energy*, vol. 329, pp. 120272, Jan. 2023.

[21] Y. Hao, Q. Lu, X. Wang and B. Jiang, "Adaptive model-based reinforcement learning for fast charging optimization of lithium-ion batteries," *IEEE Trans. Ind. Informat.*, doi: 10.1109/TII.2023.3257299.

[22] J. Wu, Z. Wei, K. Liu, Z. Quan, and Y. Li, "Battery-involved energy management for hybrid electric bus based on expert-assistance deep deterministic policy gradient algorithm," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12786–12796, Nov. 2020

[23] J. Wu, Z. Wei, W. Li, Y. Wang, Y. Li, and D. U. Sauer, "Battery thermal- and health-constrained energy management for hybrid electric bus based on soft actor-critic DRL algorithm," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 3751–3761, Jun. 2021.

[24] G. Huang, *et al.* "Real-Time Battery Thermal Management for Electric Vehicles Based on Deep Reinforcement Learning," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 14060-14072, Aug. 2022.

[25] H. Perez, X. Hu, S. Dey, and S. Moura, "Optimal charging of Li-ion batteries with coupled electro-thermal-aging dynamics," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 7761–7770, Sep. 2017.

[26] Z. Liu, Y. Wang, J. Zhang and Z. Liu, "Shortcut computation for the thermal management of a large air-cooled battery pack", *Appl. Thermal Eng.*, vol. 66, pp. 445-552, May. 2014.

[27] Z. Liu, Y. Wang, J. Zhang and Z. Liu, "Shortcut computation for the thermal management of a large air-cooled battery pack," *Appl. Thermal Eng.*, vol. 66, pp. 445-552, May 2014.

[28] C. Szepesvári and M. L. Littman, "A unified analysis of value-functionbased reinforcement-learning algorithms," *Neural Comput.*, vol. 11, no. 8, pp. 2017–2060, 1999.

[29] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, no. 11, pp. 1039–1069, 2003.

[30] J. C. Forman, S. J. Moura, J. L. Stein, and H. K. Fathy, "Genetic identification and Fisher identifiability analysis of the Doyle–Fuller–Newman model from experimental cycling of a LiFePO4 cell," *J. Power Sources*, vol. 210, pp. 263–275, Jul. 2012.

[31] S. B. Lee and S. Onori, "A robust and sleek electrochemical battery model implementation: A MATLAB® framework," *J. Electrochem. Soc.*, vol. 168, no. 9, Sep. 2021, Art. no. 090527.

[32] Yang, Moucun, et al., "Performance management of EV battery coupled with latent heat jacket at cell level." *J. Power Sources*, vol. 558, pp.232618-23229, Feb. 2023.