Electric vehicle battery charging with safe-RL

Maximiliano Trimboli[1,2], Luis Avila[1,2], and Nicolás Antonelli[1]

[1] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional, CONICET-UNSL, Av. Ejército de los Andes 950, D5700HHW San Luis, Argentina
{mdtrimboli@unsl.edu.ar, loavila@unsl.edu.ar}
[2] Facultad de Ingeniería y Ciencias Agropecuarias, Universidad Nacional de San Luis, Ruta Provincial Nº 55, D5730EKQ, Villa Mercedes, San Luis, Argentina

Abstract. To become the standard power supply for electric vehicles (EVs), Li-ion batteries need balanced current profiles in order to avoid undesirable electrochemical reactions and excessive charging times. In this work, we propose a safe exploration deep reinforcement learning (SDRL) approach in order to determine optimal charging profiles under variable operating conditions. One of the main advantages of reinforcement learning (RL) techniques is that they can learn from interaction with the real or simulated system while incorporating the nonlinearity and uncertainty derived from fluctuating environmental conditions. However, since RL techniques have to explore undesirable states before obtaining an optimal policy, no safety guarantees are provided. The proposed approach aims at maintaining zero constraint violations throughout the learning process by incorporating a safety layer that corrects the action if a constraint is likely to be violated. Tests performed on the equivalent circuit of a li-ion battery under variability conditions show early results where SDRL is able to find safe policies while considering a trade-off between the charging speed and the battery lifespan.

Keywords: Safe-RL · State of Charge · Battery aging · Variability

1 Introduction

While EVs are gaining popularity fast [11], before becoming a massive use technology of li-ion batteries must deal with two main challenges: the inconvenience of end-users and the aging of the battery [8]. Even at the fastest charging stations, it takes a considerable time to fully charge an EV. Basically, a low C-rate power supply is a limiting condition that lengthens the charging time. Meanwhile, battery aging portrays a chemical degradation mechanism that causes capacity loss and a resistance increase over the lifetime that is measured through its state of health. To diminish undesirable electrochemical side reactions within the battery, the charging process must be done with caution as aggressive current profiles can lead to severe degradation effects. At high charging speeds, li-ion batteries are prone to overheat causing them to degrade over a number of cycles which is manifested by the State-of-Health (SOH). Because of the problems with fast

charging, EV batteries have built-in charging speed limits set by battery management systems. It's clear the necessity of compromises between the battery core temperature and time of charge [19], and thus the algorithms determining the charging profiles will have a strong influence on the final performance. A manner to face this trade-off is to incorporate constraints on the battery model used to learn new charging polices. However, as the behavior of a li-ion battery is difficult to predict because nonlinearities, variability and its dependence on fluctuating environmental conditions [5, 9], a precise model results expensive to be determined. Reinforcement Learning (RL) techniques allow to obtain an optimal control policy without relying on a detailed model of the system being controlled.

A number of works have employed RL to tackle the problem of finding an optimal charging policy. [4] proposes a battery charging control methodology based on RL to minimize the charging costs. In [17] the author uses an RL-based dynamic charging methodology with multiple active modes to address the problem of extending the battery lifetime. In [12] an optimal charging strategy considering the battery life extension based on RL is presented. An adaptive charging technique for li-ion battery using RL and multi-stage-constant-current is presented in [18]. Two methods, described as selective and generic policy approaches, are considered to train the algorithm and further deployed for a range of initial SOCs. Finally, [15] presented a fast-charging strategy based on a gradient-policy actor-critic framework for li-ion batteries.

Safe-RL is a topic that grows in relevance and considers learning problems in which it is crucial that the agent interacts with the environment only through safe policies, i.e., policies that do not take the agent to undesirable situations [6]. Since RL focuses on maximizing the long-term reward, it is likely to visit undesirable states during the learning process if no safety guarantees are provided. This may lead real systems to a failure or harmful condition before an optimal policy can be learned. A novel approach is to directly add a safety signal to the learned policy to perform a correction over the action towards safety limits [7]. One advantage of the technique is that it provides a closed-form solution through a linearized model learned from past trajectories generated under random actions.

In this work, we propose this approach to the problem of finding fast-charging strategies regarding the battery core temperature. This manner, the goal of maintaining zero-constraint violations throughout the learning process relies on a pre-trained neural network that predicts the change in a safety signal over a single time step. The trained model is incorporated into a safety layer that corrects the action if a constraint is likely to be violated. The proposed method is tested in the equivalent circuit of a li-ion battery considering variability conditions.

The article is structured as follows: Section 2 provides an overview of RL with safe exploration for continuous spaces. In Section 3, a brief explanation of the operation of the environment and the results obtained for optimal battery charge are shown and discussed. Finally, in Section 4 some remarks and future research efforts are highlighted.

2 Li-ion battery model

Lithium-iron-phosphate is the predominant rechargeable battery chemistry used for EVs for their superior volumetric energy density and charge efficiency. To maintain certain balance between accuracy and complexity, an equivalent-circuit electrical model integrated with a two-state thermal model is used to describe the li-ion battery dynamic, as shown in Fig. 1. Both the system of equations that define the operation of the model and the selection of parameters have been selected and adjusted according to [16]. The thermal model is incorporated since the electrical parameters depend on the battery core temperature resulting dynamics is highly nonlinear and it can be higher than the surface temperature under high current rates [14].

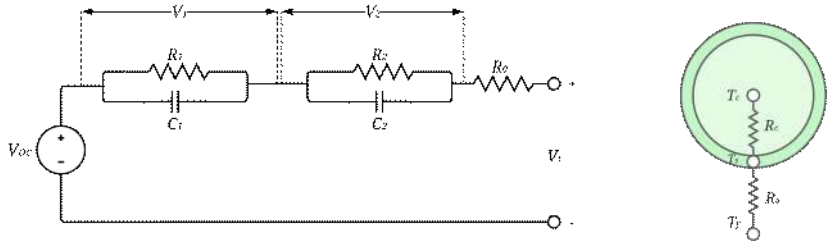

Fig. 1. Equivalent circuit of the electro-thermal model

2.1 Equivalent-circuit model with battery aging

The electrical subsystem consists of an equivalent circuit formed by an open-circuit voltage source (OCV, V_{oc}), two RC pairs (OCV, V_{oc}) and a resistor (R_0) connected in series. While the thermal subsystem consists of a model that describes the radial heat transfer dynamics of a cylindrical battery by considering the dynamics of both core and surface temperatures T_c and T_s respectively.

In this proposed system, we consider charging current and ambient temperature as inputs and voltage as the measured output is used for describing the battery behavior.

Besides, the aging subsystem is based upon a matrix of cycling tests from [3]. The semi-empirical life model adopted the following equation to express the correlation between the capacity loss (ΔQ_b, in %) and the discharged A_h throughput. The end-of-life (EOL) of an automotive battery is defined by a capacity loss of 20%.

The SOH provides a form of implement the battery aging model from its time derivative:

$$\frac{dSOH}{dt} = -\frac{|I(t)|}{2N(c, T_c)\, C_{bat}} \tag{1}$$

where $I(t)$ corresponds to the instant charging current, N is the number of cycles, c represents the C-rate and C_{bat} indicates the nominal capacity of the battery. The equation shows an explicit dependence between the battery aging process with the core temperature and the charging current profile employed.

2.2 Variability in the state of the battery

Operating conditions change with temperature when they are subject to wide temperature variations. Temperature-dependent characteristics derived from design and aging, prevent accurately determining the battery's remaining capacity and its health condition. Furthermore, such variability is a burden for current protective policies.

A practical yet simple alternative to describe battery fluctuating behavior due to temperature is using a stochastic diffusion process. Ito [10] provided an alternative to ordinary numerical rules of calculus by defining a particular kind of uncertainty representation based on a Wiener diffusion process. the Ito's process parameter is included in the deterministic thermal model that describes the core temperature T_c, such that it affects with a given probability the percentage of available charge remaining in the battery.

This mathematical model ensures a manifold of alternative dynamic behaviors that accounts sufficiently well for the intra-variability. The aim of including a stochastic process is to obtain an analytical representation of the potential variability among real battery cells. It is worth noting that ambient temperature, as well as battery hyperparameters, remain constant along the simulation and only the Ito parameter causes the SOC fluctuation. Notice that lithium batteries are not expected to behave in terms of a stochastic process but, by representing the battery dynamics in this way, we ensure the SOC curve reaches all possible states for different operating condition. Eventually, accounting for variability in the SOC estimation can help to minimize the risk of premature failure caused by over-charging and over-discharging events and provide information for the management system to keep the battery working within a safe operating window.

3 Safe Reinforcement Learning

3.1 Constrained MDP

We study a special case of constrained Markov decision processes (CMDP) [1] where the observed safety signals should be kept bounded. A CMDP is characterized by the tuple (S, A, P, R, γ, C), where S is a state space, A is an action space, $P : S \times A \times S \to [0; 1]$ is a transition kernel, $R : S \times A \to R$ is a reward function, $\gamma \in (0; 1)$ is the discount factor, and $C = \{c_i : S \times A \to R \,|\, i \in [K]\}$ is

a set of immediate-constraint functions, given the set K formed by $\{1, \ldots, K\}$. Based on that, we also define a set of safety signals $\bar{C} = \{\bar{c}_i : S \to R \,|\, i \in [K]\}$ as per-state observations of the immediate-constraint values. Finally, let policy $\mu : S \to A$ be a stationary mapping from states to actions.

We therefore study safe exploration in the context of policy optimization, where at each state, all safety signals $\bar{c}_i(\cdot)$ are upper bounded by corresponding constants $C_i \in R$:

$$\max_{\theta} \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t R\left(s_t, \mu_\theta\left(s_t\right)\right)\right] \text{s.t.} \quad \bar{c}_i\left(s_t\right) \leq C_i \; \forall i \in [K] \tag{2}$$

where μ_θ is the parameterized policy.

3.2 Linear Safety-Signal Model

Solving Eq. (2) is a hard task, even for a simple model describing the battery dynamics. A key aspect to this challenge is the RL agent's intrinsic need to explore to find new and improved actions. However, it is not possible ensure per-state constraint satisfaction at early training stages using a random policy without prior knowledge on its environment. Notice this statement is still true when the reward is carefully shaped to penalize unsafe states as for an RL agent to learn to avoid undesired behavior it will have to violate the constraints enough times for the negative effect to propagate in the dynamic programming scheme.

To mitigate this problem, we can add some basic form of prior knowledge based on single-step dynamics. Single-step transition data in logs is rather common and more realistic compared to also knowing behavior policies. We do not attempt a learning of the full transition model, but solely the immediate-constraint functions $c_i(s, a)$. Considering $[x]^+$ as the operation $\max\{x, 0\}$; where $x \in R$, in (15) we perform a linearization to obtain a first-order approximation to $c_i(s, a)$ with respect to a.

$$\bar{c}_i\left(s\right) \triangleq c_i\left(s, a\right) \approx \bar{c}_i\left(s\right) + g(s, w_i)^\top a \tag{3}$$

where w_i are weights of a neural network $g(s; w_i)$, taking s as input and outputs a vector of the same dimension as a. This model is an explicit representation of sensitivity of changes in the safety signal to the action using features of the state.

From a set of tuples $D = \left\{\left(s_j, a_j, s'_j\right)\right\}$ independent of policy, we train $g(s; w_i)$ by solving

$$\arg\min_{w_i} \sum_{(s,a,s') \in D} \left(\bar{c}_i\left(s'\right) - \left(\bar{c}_i\left(s\right) + g(s, w_i)^T a\right)\right)^2 \tag{4}$$

were D is generated initializing the agent in a uniformly random location to perform actions of similar characteristics along multiple episodes that end when a time interval expires or when a constraint violation occurs. Training $g(s; w_i)$

on D is performed once per task as a pretraining phase that precedes the RL training.

3.3 Safety Layer via analytical Optimization

To solve problem Eq. (2) we using Deep Deterministic Policy Gradient (DDPG) [13], a policy gradient algorithm [2] whose policy network directly outputs actions and not their probabilities.

Using the deterministic action $\mu_\theta(s)$ selected by the deep policy network, we use an additional layer located on top of the policy network, whose function is to solve

$$a^* = \arg\min_a 12\|a - \mu_\vartheta(s)\|^2 \text{s.t.} \quad \bar{c}_i\left(s_t\right) + g\left(s, w_i\right)^\top a \leq C_i \forall i \in [K] \tag{5}$$

Based on the assumption that no more than one constraint is active at the same time we can gain the benefit of obtaining a closed-form analytical solution to Eq. (5). Therefore, assuming the existence of a closed-form solution to Eq. (5) denoted by $(a^*, \{\lambda_i^*\}_{i=1}^K)$, where λ_i^* is the optimal Lagrange multiplier associated with the i-th constraint, and that $|\{i|\lambda_i^* > 0\}| \leq 1$; i.e., at most one of the constraints is active, then

$$\lambda_i^* = \left[\frac{g(s; w_i)^\top \mu_0(s) | \bar{c}_i(s) - C_i}{g(s; w_i)^\top g(s; w_i)}\right]^+ \tag{6}$$

and

$$a^* = \mu_\theta(s) - \lambda_{i^*}^* g(s; w_{i^*}) \tag{7}$$

where $i^* = \arg\max_i \lambda_i^*$.

The solution Eq.(7) is basically a linear projection of the original action $\mu_\theta(s)$ to the "safe" hyperplane with slope $g(s; w_i^+)$ and intercept $\bar{c}_{i^*}(s) - C_{i^*}$.

4 Experiments

In this section, we evaluated the proposed method in the simulated environment of the li-ion battery model and we compared the results obtained using the proposed methodology with other approaches such as reward shaping.

4.1 Experimental Setup

We assume the following parameters for the li-ion battery model: a nominal cell capacity $C_{bat} = 2300Ah$, with a maximum charging current $I(t) = 46A$ and a minimal terminal voltage $V_{t_{min}} = 2V$. Initially, the core T_c and surface T_s battery temperatures are at the same as the ambient temperature T_f.

Being the temperature an important variable in order to extend the life cycle of a li-ion battery, we included the mean temperature T_m term in the continuous state-space model. The charging current corresponds to the action chosen by the agent, over a continuous action space, and applies to the environment as shown in Fig. 2.

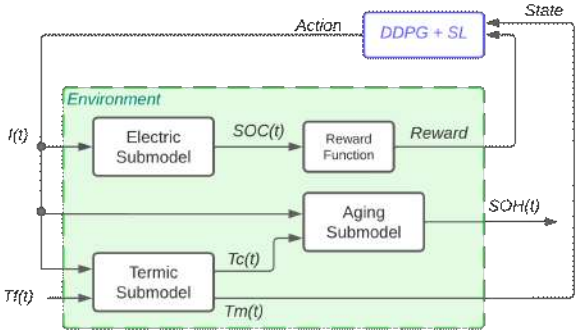

Fig. 2. Simplified agent-environment interaction scheme

The negative sign in the variable corresponds to a charging current. The range limits for state and action spaces are defined by safe operating limits for a li-ion battery and are specified in Table 1, along with other parameters of the environment. The state and the action are normalized to the range $[0, 1]$ so as to increase stability during training of the algorithms. All the episodes were initialized under the same environmental conditions for ease of comparison.

To avoid excessive manipulation during learning and to appreciate the real effect of the safety layer over the obtained policy, we designed the following reward function that depends on the actual SOC value:

$$R = \begin{cases} SOC - 1, & \text{if } 0 < SOC < 1 \\ -1, & \text{otherwise} \end{cases} \tag{8}$$

The hyper-parameters used for training are shown in Table 2. Target networks are composed of the same structure as their corresponding peer (actor or critic network).

Agent learning consists of many training epochs, each of these followed by an evaluation phase after a fixed number of steps. Also, for each epoch there is a variable number of episodes used to update the buffer memory and reset the environment conditions. The episode ends when a maximum step length is reached or the agent reaches the maximum SOC. To obtain the safety signals c_i, a pre-training phase is performed to compute the Lagrange multipliers and the action correction term. The remaining parameters used to describe the safety model can be observed in Table 3.

Table 1. Main hyperparameters of battery environment

Hyperparameter	Symbol	Ranges	Unit
State Space	S	[5, 45]	$^{\circ}C$
Action Space	A	[-46, 0]	A
Initial SOC	SOC_0	0.3	
Initial SOH	SOH_0	0.9	
Ambient Temperature	T_f	25	$^{\circ}C$

Table 2. Main hyperparameters of DDPG algorithm

Hyperparameter	Symbol	Ranges
Actor layers	σ	[128 64]
Critic layers	ϕ	[64, 128, 32]
Epochs	e_{DDPG}	100
Steps per epoch at training	m_t	6000
Steps per epoch at evaluation	m_e	1500
Maximum episode length	l_{max}	300
Batch size	B	256
Replay Buffer Size	D	1000000
Discount factor	γ	0.99
Actor learning rate	α_σ	0.00001
Critic learning rate	α_ρ	0.0001
Action Noise Range	ϵ	0.01
Optimizer		Adam

Table 3. Main hyperparameters of safety layer module

Hyperparameter	Symbol	Ranges
Constraint model layers	C	[2, 2]
Epochs	e_{SL}	5
Steps per epoch at training	m_t	6000
Steps per epoch at evaluation	m_e	1500
Maximum episode length	l_{max}	300
Batch size	B	256
Learning rate	α_{SL}	0.001

4.2 Experimental Results

To compare the effectiveness of the proposed method we implemented a DDPG technique without constraints and a DDPG strategy with reward shaping. Reward shaping manipulates the reward signal to provide the agent with expert knowledge to avoid undesired areas. Specifically, the agent is penalized with a negative reward when the temperature T_m surpasses the temperature margin M. The margin M is the upper limit at which the policy begins to correct its actions so that the temperature does not exceed the restrictions. To determine the best value for M, we set the penalty to $r = -1$ when the agent is above M and rolled out series of 10 simulations with different seeds for $M \in \{0.05, 0.09, 0.12, 0.15\}$.

The accumulated constraint violations, i.e. times the agent is outside the safety bounds, for each run was determined.

The box plot in Fig. 3 represents at what extent the reward shaping strategy is able to avoid the safety limits for each value M. As observed, for a margin $M = 0.12$ the strategy shows the best performance. However, considering that as the safety margin grows the battery capacity to reach the maximum SOC decreases, we also consider a lower value $M = 0.05$ in the experiments. Notice that as the margin grows, the number of violations goes down. This implies a longer time to charge the battery. Therefore, the margin implies a trade-off between the charging time and the number of safety breaks.

Fig. 4 shows the accumulated rewards obtained with the implemented strategies during training phase. Even though all of them converge to the respective policy, we can observe that using a shaped reward function causes more instability during training, with large negative peaks of negative rewards produced by constraints violations. The best performance is obtained by the DDPG strategy without any restrictions. Unrestricted DDPG and DDPG+RS with $M = 0.05$ are severely penalized at the beginning of the training process due to the excessive number of constraint violations. DDPG+RS with $M = 0.12$ achieves less reward as it spends more time to get $SOC = 1$.

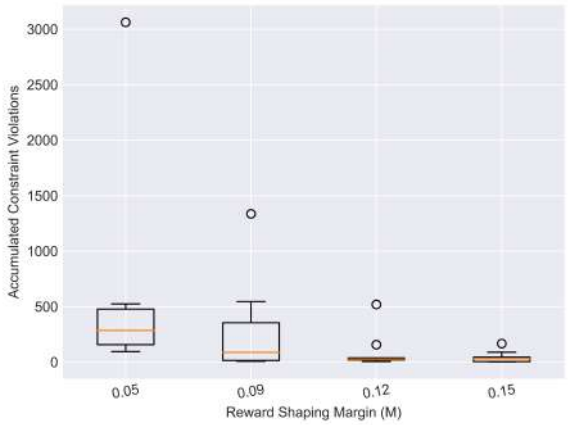

Fig. 3. Box plot of accumulated constraint throughout the training of DDPG with reward shaping

We also obtained relevant metrics during the learning phase, which are shown in Table 4. We compared for each technique the number of episodes and the average number of steps per episode to converge. The number of violations accounts for the percentage of total episodes in which the temperature exceeds the limit. Looking at the average number of steps required to complete a battery charge,

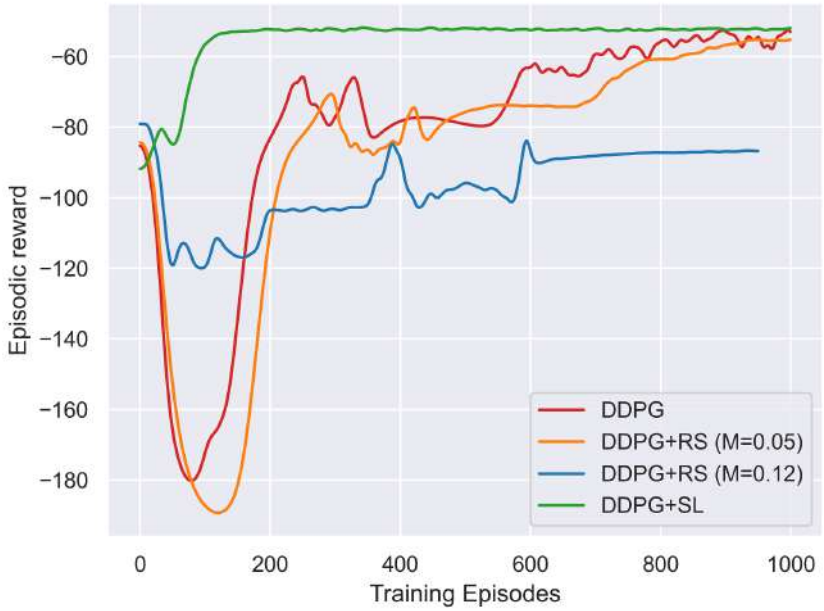

Fig. 4. Cumulative reward through learning process obtained

we can see that DDPG charges the battery faster at the expense of a higher number of constraint violations. In turn, the SDRL strategy allows the battery to charge slightly faster than the RS methods without violating any constraints. As stated, lithium ion batteries are prone to overheating during high charging current rates, causing them to degrade over multiple cycles. We can see in the last row the resulting SOH after the training process. SDRL far outperforms all other algorithms, DDPG+RS with $M = 0.12$ reaches a %40 of degradation (taking account a $SOH_0 = 0.9$ according to Table 1) but with longer charging time while unrestricted DDPG achieves %50 of SOH degradation.

Figures 5-8 show the obtained curves for different charging policies. Different behaviors can be observed depending on the temperature limitation. The DDPG algorithm (Fig. 5) charges the battery as quickly as possible using the maximum current allowed, violating the restrictions. Because the policy does not account for temperature, the current is nearly constant, and $SOC = 1$ is reached around Step 150. DDPG+RS with $M = 0.05$, also uses the maximum allowed current to charge the battery until the temperature reaches the margin, as in Fig. 6. The policy then limits the current and increases the number of steps required to achieve $SOC = 1$ to 200. Since DDPG+RS with $M = 0.12$ uses a larger margin, as shown in Fig. 7, it starts correcting the current profile sooner. At the expense of a large number of steps to charge the battery, we can observe

a constant current profile with no restriction violations. The proposed SDRL strategy operates near the temperature restriction but without any violations, as observed in Fig. 8, while the time of charge is between the unrestricted DDPG policy and those values achieved with RS.

Table 4. Battery charging data after training

	DDPG	DDPG+RS ($M = 0.05$)	DDPG+RS ($M = 0.12$)	DDPG+SL
Avg. steps for full charge	**154.25**	205.83	251	174.43
% of ep. with restriction violations	30.96%	9.09%	1.52%	**0%**
SOH value at convergence	0.3978	0.4872	0.5304	**0.8602**

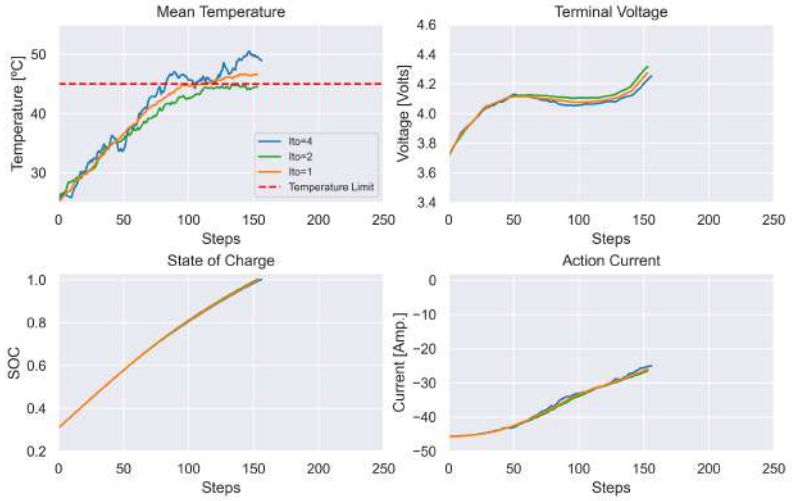

Fig. 5. Battery charging process using only DDPG algorithm

5 Concluding remarks

One daunting factor for gas-powered vehicle owners considering switching to EVs is the time it takes to charge the battery. Aggressive charge profiles can reduce waiting time, but produce undesired electrochemical effects. It is thus necessary to find charging profiles that keep the battery in an optimal range to maximize battery life. In this paper, we use a SDRL approach to obtain high-quality charging profiles for a li-ion battery using RL algorithms that never violate constraints

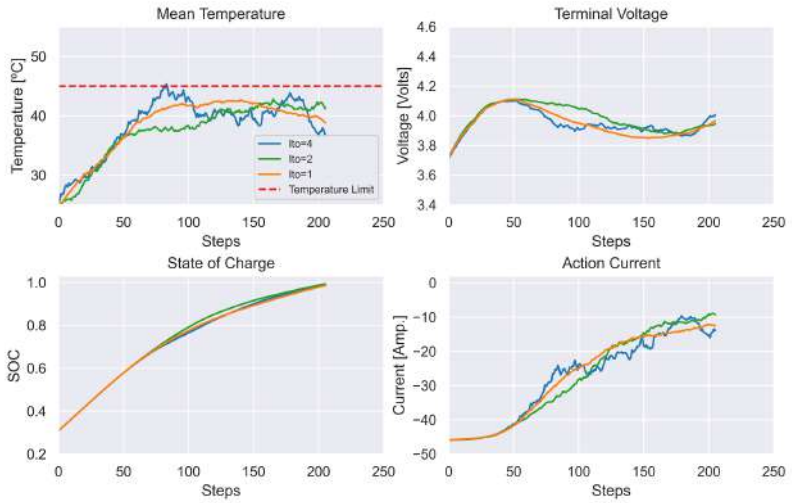

Fig. 6. Battery charging process using DDPG+RS with margin $M = 0.05$

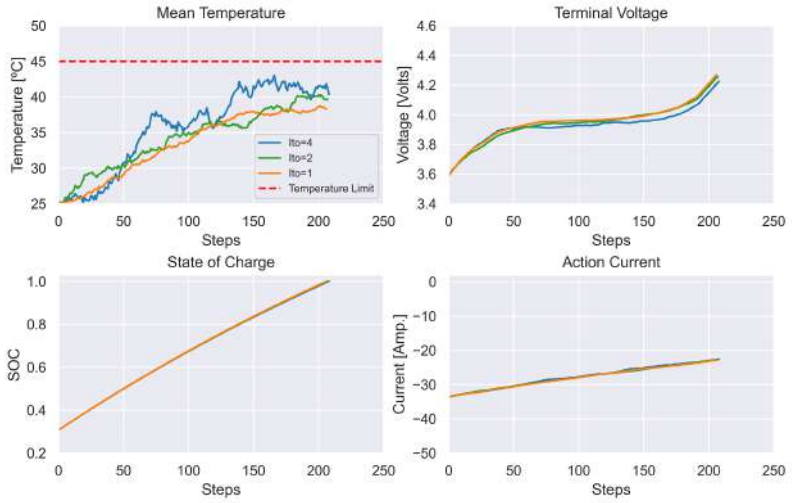

Fig. 7. Battery charging process using DDPG+RS with margin $M = 0.12$

during learning. Specifically, we propose a DDPG algorithm to compute the policy and a safety layer that analytically solves an action correction formulation per each state. This technique provides an elegant closed-form solution through a linearized model learned on past trajectories consisting of random actions. We

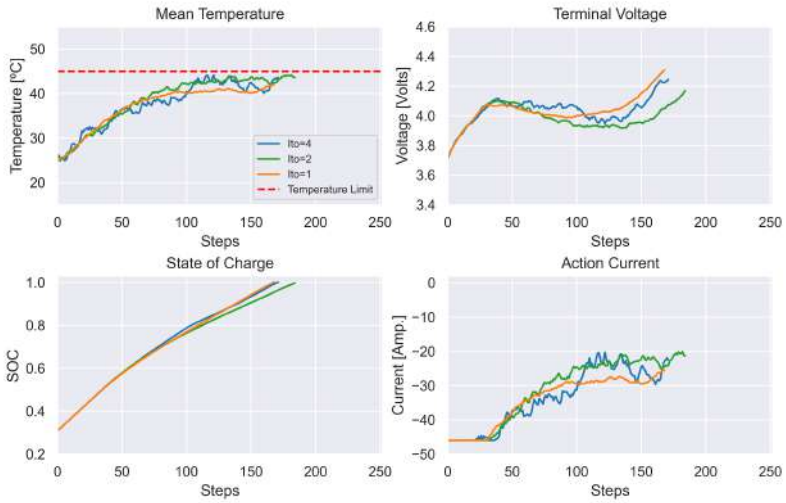

Fig. 8. Battery charging process using DDPG algorithm with Safety Layer

performed a number of experiments on an equivalent circuit that simulates the battery dynamics under variable operating conditions and compared the results with benchmark methods. Since the RL approach used is model-free, it learns through direct interactions with the controlled system regardless of complexity or parameterization. In general terms, the proposed model was more efficient considering characteristics like time to complete a charge, smoothness of the current profile and battery life maintenance.

References

1. Altman, E.: Constrained Markov decision processes, vol. 7. CRC Press (1999)
2. Baxter, J., Bartlett, P.L.: Infinite-horizon policy-gradient estimation. Journal of Artificial Intelligence Research **15**, 319–350 (2001)
3. Campbell, I.D., Gopalakrishnan, K., Marinescu, M., Torchio, M., Offer, G.J., Raimondo, D.: Optimising lithium-ion cell design for plug-in hybrid and battery electric vehicles. Journal of Energy Storage **22**, 228–238 (apr 2019). https://doi.org/10.1016/J.EST.2019.01.006, https://www.sciencedirect.com/science/article/pii/S2352152X18300094
4. Chang, F., Chen, T., Su, W., Alsafasfeh, Q.: Control of battery charging based on reinforcement learning and long short-term memory networks. Computers & Electrical Engineering **85**, 106670 (2020)
5. Cho, S., Jeong, H., Han, C., Jin, S., Lim, J.H., Oh, J.: State-of-charge estimation for lithium-ion batteries under various operating conditions using an equivalent circuit model. Computers Chemical Engineering **41**, 1–9 (jun 2012). https://doi.org/10.1016/J.COMPCHEMENG.2012.02.003, https://www.sciencedirect.com/science/article/pii/S0098135412000464

6. Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., Ghavamzadeh, M.: Lyapunov-based Safe Policy Optimization for Continuous Control (jan 2019), http://arxiv.org/abs/1901.10031
7. Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., Tassa, Y.: Safe Exploration in Continuous Action Spaces (2018), http://arxiv.org/abs/1801.08757
8. Danilov, D., Notten, P.H.: Adaptive battery management systems for the new generation of electrical vehicles. In: 5th IEEE Vehicle Power and Propulsion Conference, VPPC '09. pp. 317–320 (2009). https://doi.org/10.1109/VPPC.2009.5289835
9. Dubarry, M., Pastor-Fernández, C., Baure, G., Yu, T.F., Widanage, W.D., Marco, J.: Battery energy storage system modeling: Investigation of intrinsic cell-to-cell variations. Journal of Energy Storage **23**, 19–28 (jun 2019). https://doi.org/10.1016/J.EST.2019.02.016, https://www.sciencedirect.com/science/article/pii/S2352152X18308156
10. Ito, K.: Stochastic differentials. Applied Mathematics Optimization **1**, 374–381 (1975)
11. Jordán, J., Palanca, J., Martí, P., Julian, V.: Electric vehicle charging stations emplacement using genetic algorithms and agent-based simulation. Expert Systems with Applications **197**, 116739 (2022)
12. Kim, M., Baek, J., Han, S.: Optimal Charging Method for Effective Li-ion Battery Life Extension Based on Reinforcement Learning (may 2020), http://arxiv.org/abs/2005.08770
13. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)
14. Lin, X., Perez, H.E., Mohan, S., Siegel, J.B., Stefanopoulou, A.G., Ding, Y., Castanier, M.P.: A lumped-parameter electro-thermal model for cylindrical batteries. Journal of Power Sources **257**, 1–11 (jul 2014). https://doi.org/10.1016/J.JPOWSOUR.2014.01.097, https://www.sciencedirect.com/science/article/abs/pii/S0378775314001244
15. Park, S., Pozzi, A., Whitmeyer, M., Perez, H., Joe, W.T., Raimondo, D.M., Moura, S.: Reinforcement Learning-based Fast Charging Control Strategy for Li-ion Batteries. CCTA 2020 - 4th IEEE Conference on Control Technology and Applications pp. 100–107 (feb 2020), http://arxiv.org/abs/2002.02060
16. Perez, H.E., Hu, X., Dey, S., Moura, S.J.: Optimal Charging of Li-Ion Batteries With Coupled Electro-Thermal-Aging Dynamics. IEEE Transactions on Vehicular Technology **66**(9), 7761–7770 (sep 2017). https://doi.org/10.1109/TVT.2017.2676044, http://ieeexplore.ieee.org/document/7867072/
17. Triki, M., Ammari, A.C., Wang, Y., Pedram, M.: Reinforcement learning-based dynamic power management of a battery-powered system supplying multiple active modes. In: Proceedings - UKSim-AMSS 7th European Modelling Symposium on Computer Modelling and Simulation, EMS 2013. pp. 437–442. IEEE Computer Society (2013). https://doi.org/10.1109/EMS.2013.74
18. Tunuguntla, S.T.: Adaptive charging techniques for Li-ion battery using Reinforcement Learning (2020), https://etd.iisc.ac.in/handle/2005/5032
19. Xing, Y., Ma, E.W.M., Tsui, K.L., Pecht, M., Xing, Y., Ma, E.W.M., Tsui, K.L., Pecht, M.: Battery Management Systems in Electric and Hybrid Vehicles. Energies **4**(11), 1840–1857 (oct 2011). https://doi.org/10.3390/en4111840, http://www.mdpi.com/1996-1073/4/11/1840