Digital Object Identifier 10.1109/PEDG54999.2022.9923188

2022 IEEE 13th International Symposium on Power Electronics for Distributed Generation Systems (PEDG)

Reinforcement Learning Based Modulation for Balancing Capacitor Voltage and Thermal Stress to Enhance Current Capability of MMCs

Jun-Hyung Jung

Ehsan Hosseini

Marco Liserre

Luis M. Fernández-Ramírez

Suggested Citation

J. -H. Jung, E. Hosseini, M. Liserre and L. M. Fernández-Ramírez, "Reinforcement Learning Based Modulation for Balancing Capacitor Voltage and Thermal Stress to Enhance Current Capability of MMCs," 2022 IEEE 13th International Symposium on Power Electronics for Distributed Generation Systems (PEDG), 2022, pp. 1-6, doi: 10.1109/PEDG54999.2022.9923188.

Reinforcement Learning Based Modulation for Balancing Capacitor Voltage and Thermal Stress to Enhance Current Capability of MMCs

Jun-Hyung Jung
Chair of Power Electronics
Kiel University
Kiel, Germany
jj@tf.uni-kiel.de

Ehsan Hosseini
Research Group in Sustainable Technologies
University of Cadiz
Algeciras, Spain
ehsan.hosseini@alum.uca.es

Marco Liserre
Chair of Power Electronics
Kiel University
Kiel, Germany
ml@tf.uni-kiel.de

Luis M. Fernández-Ramírez
Research Group in Sustainable and Renewable Electrical Technologies
University of Cadiz
Algeciras, Spain
luis.fernandez@uca.es

***Abstract*— Balancing DC capacitor voltage of many submodules (SMs) is one of the important issues in modular multilevel converter (MMC) systems. In addition, the balance of thermal stress between SMs should be considered to equalize the lifetime expectation of semiconductors and to enhance the current capability of MMC systems. However, it is complicated to balance all the various factors satisfactorily at the same time. Recent machine learning (ML) techniques can achieve optimal results through learning using numerous data acquired in complex environments. Therefore, this paper proposes a new modulation based on reinforcement learning (RL), which is a subclass of ML methods, to optimally balance the capacitor voltage and thermal stress of SMs. A deep Q-network (DQN) agent, which is one of the RL algorithms, is applied in accordance with a nearest-level modulation (NLM), and main features of the DQN agent are described in this paper. The effectiveness of the proposed modulation based on RL is verified by simulations results.**

***Keywords*— Modular multilevel converter, submodule DC capacitor voltage, thermal balancing, reinforcement.**

I. INTRODUCTION

A MMC is a promising topology for high-voltage and medium-voltage applications because of easy DC voltage scalability by connecting SMs in series [1]. Among modulations used for MMC, phase-shift PWM (PSPWM) and NLM are mainly used, and NLM is more suitable for MMC systems with a large number of levels because it is possible to output the current close to a sinusoidal waveform without using PWM, which leads to significant switching losses [2, 3]. However, the amount of change in the SM DC capacitor voltage is large due to the long continuous conduction time of SMs. Therefore, various studies have been conducted the balancing control of the SM DC voltage with sorting methods [4, 5].

In addition to the SM voltage balancing control, the importance of equalizing thermal stress between semiconductors has emerged recently in several studies [6–8]. The imbalance of the thermal stress not only reduces the current capability of converters [7] but also increases design costs due to additional margin for the safe operation. Furthermore, different lifetime expectancy of semiconductors deteriorates reliability and safety of the converter system. To distribute the thermal stress evenly, cost functions considering semiconductor losses and junction temperature are introduced in [7] and [8].

However, satisfying optimal results for multi-purposes under different conditions, e.g. balance of the thermal stress and the SM DC voltage, requires very complex control systems and algorithms. Lately, research on RL, which is a subclass of ML, that can provide a solution to control complex and uncertain systems is receiving a lot of attention [9]. In RL, an agent, a learner including RL algorithm and deep neural networks, can be trained by direct interaction with the environment, such as PWM converters, to predict and control the system optimally. In power electronics applications, many papers have proposed RL based modulations and control systems to achieve optimal results [10]. In [11], optimal phase shift angles for reducing the power dissipation caused in dual-active-bridge converter is reached through a Q-learning based modulation. In [12], a DQN agent based RL is used to stabilize DC bus voltage of a buck DC-DC converter. The RL based controller can smooth the voltage tracking while operating under large power loads variations. The paper [13] proposes a space vector modulation based on RL using the DQN agent for d-q current and neutral-point voltage controls in a three-level NPC converter. In spite of effective results, the use of RL for simple tasks that can be solved with existing control algorithms should be reconsidered. In [14], a study on the application of RL with policy-gradient algorithm is also conducted for MMCs, which is relatively more complicated than other converters, to control phase current, circulating current and SM DC capacitor voltage balance. However, it is not appropriate applying RL for the entire control system rather than specific purposes, because of increase in training time and computational burden.

This paper proposes a new modulation technique based on RL for balancing SM DC capacitor voltage and thermal stress on semiconductors in MMC systems. In order to obtain a switching state satisfying optimally balanced conditions, the DQN agent, a representative RL algorithm, is applied to the modulation, and the states used for training are defined to balance the DC capacitor voltage and semiconductor losses. In addition, hyperparameters and reward functions used for training are also described. The effectiveness of the proposed RL-based modulation is verified in MATLAB simulation.

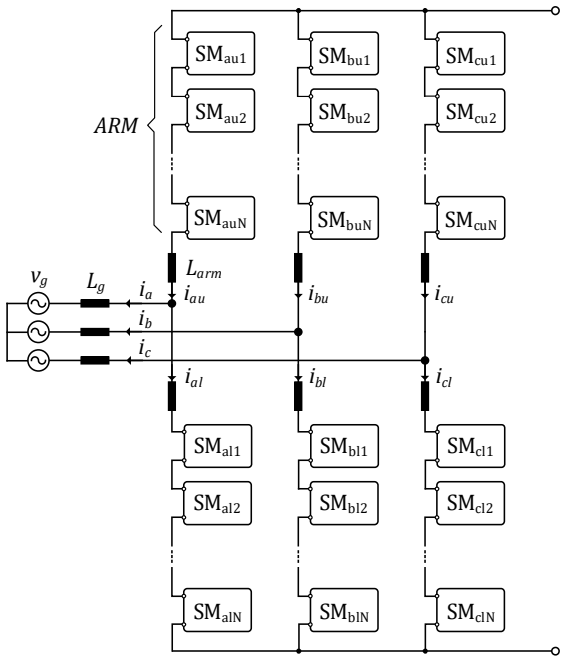


Fig. 1. Three-phase N-level MMC system.

This paper is organized as follows. Section II describes a basis of MMC systems using the NLM modulation and DC capacitor voltage and junction temperature imbalance between SMs. In Section III, RL and DQN agent are introduced first and the how the DQN agent is applied and trained to balance the SM DC capacitor voltage and junction temperature of semiconductors. Section IV shows simulation conditions and results for the validation of the performance of the proposed modulation based on RL. Finally, Section V draws the conclusions of this paper.

II. Imbalance of SM DC Capacitor Voltage and Junction Temperature in MMC

A. NLM in MMC

Fig. 1 shows the three-phase MMC system that consists of six arms, and each arm has a structure in which one inductor L_{arm} and SMs are connected in series. Each SM is generally configured with a half-bridge converter composed of four semiconductors (S_1, S_2, D_1 and D_2) and a DC capacitor (C_{SM}). The serial connection of SMs enables MMC to accommodate higher DC voltage than conventional PWM converters.

The most common three modulation methods for MMC are PSPWM, level-shift PWM (LSPWM) and NLM. The PSPWM and LSPWM are preferred in low-level MMC systems, while NLM can be used for high-level MMC systems because it has less switching loss than PWM based methods and results in good performance enough if the number of SMs used in MMC is high enough.

In a single-leg of the MMC, upper and lower arm voltage references are expressed in (1) and (2) [1]:

$$v_{Ux}^{*} = \frac{V_{DC}}{2} - v_{xn}^{*} - v_{xd}^{*} \tag{1}$$

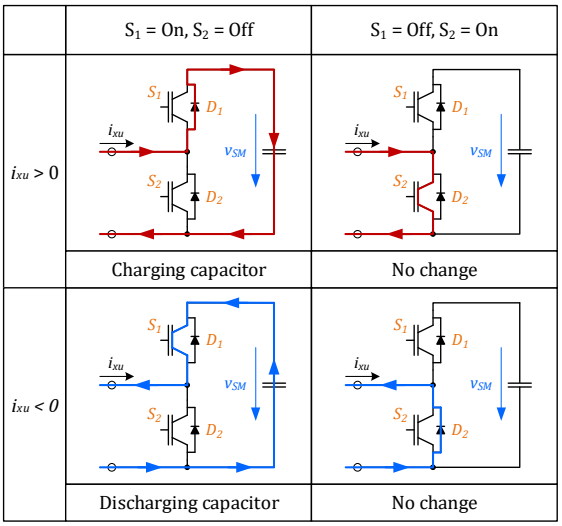


Fig. 2. Effect of SM states and current direction on DC capacitor voltage.

$$v_{Lx}^{*} = \frac{V_{DC}}{2} + v_{xn}^{*} - v_{xd}^{*} \tag{2}$$

where, v_{Ux}^{*} and v_{Lx}^{*} are upper and lower arm voltage references, v_{xn}^{*} is a voltage reference of phase x (x={a, b, c}) and v_{xd} is a voltage reference to control circulating current flowing through phase x. Assuming that the capacitor voltages of N SMs are equal to each other as V_{SM}, from (1) and (2), the nearest voltage output levels for the two arms in phase x, N_{Ux}^{*} and N_{Lx}^{*}, are as follows,

$$N_{Ux}^{*} = Round(v_{Ux}^{*}/V_{SM}) \tag{3}$$

$$N_{Lx}^{*} = Round(v_{Lx}^{*}/V_{SM}) \tag{4}$$

B. SM DC Capacitor Voltage and Junction Temperature of Semiconductors in SMs

In terms of balancing condition of many SMs, the general control issue of MMC systems is the balance control of the SM DC capacitor voltage, v_{SM_n} (n=1, 2, 3, …, N). This voltage is varied according to the switching state and the arm current flowing through the SM, as shown in Fig. 2. It can be expressed in (5):

$$v_{SM}^{N}(t) = \frac{1}{C_{SM}} \int_{0}^{T_s} i_{arm} \cdot S^{N}(t) dt + v_{SM}^{N}(0) \tag{5}$$

where $S^{N}(t)$ is a switching state of SMs and $v_{SM_n}(0)$ is the initial voltage of each period, T_s.

In terms of SM voltage, PSPWM does not cause a sudden voltage imbalance between SMs because the SM insertion times are relatively evenly distributed. However, in the case of NLM, the SM voltage imbalance easily occurs more than PSPWM because the insertion times of each SM are quite different due to a low frequency of switching state changes. To balance the SM voltage when the NLM is used for the MMCs, SM voltage balancing control methods based on the sorting algorithm is used to decide the SM to be inserted, which has the highest or lowest

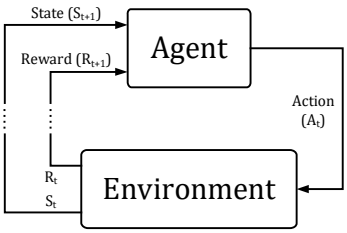


Fig. 3. Basic configuration of RL

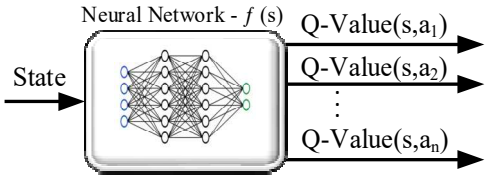


Fig. 4. State mapping to Q-values in DQN

voltage level, according to the arm current direction to generate the arm voltage. [4, 5].

In addition to the SM voltage balancing control, the thermal stress balancing of semiconductors between SMs is also important to enhance the current capability and to increase the lifetime of the MMC system [7, 8]. Junction temperature of the semiconductor (T_j) is affected by loss caused in semiconductor (P) and parameters of thermal equivalent circuits, such as a thermal resistance ($R_{th(j-a)}$) between the junction and ambient temperature (T_a), as given in (6).

$$T_j = T_a + R_{th(j-a)} \times P \tag{6}$$

The average conduction losses caused in the SM can be expressed with the insertion state of the SM and the arm current in (7)-(10). [7] The switching losses are ignored due to low switching state changes.

$$\bar{P}^n_{T1.c}(t) = \frac{1}{T_{avg}} \int_{t-T_{avg}}^{t} S^n(t) v^n_{ce}(t) \, |i_{arm}(t)| dt \tag{7}$$

$$\bar{P}^n_{T2.c}(t) = \frac{1}{T_{avg}} \int_{t-T_{avg}}^{t} \{1 - S^n(t)\} v^n_{ce}(t) \, i_{arm}(t) dt \tag{8}$$

$$\bar{P}^n_{D1.c}(t) = \frac{1}{T_{avg}} \int_{t-T_{avg}}^{t} S^n(t) v^n_{f}(t) \, i_{arm}(t) dt \tag{9}$$

$$\bar{P}^n_{D2.c}(t) = \frac{1}{T_{avg}} \int_{t-T_{avg}}^{t} \{1 - S^n(t)\} v^n_{ce}(t) \, |i_{arm}(t)| dt \tag{10}$$

Similar to the SM DC capacitor voltage imbalance, the junction temperature of semiconductors between SMs are easy to be unbalanced, even if the switching loss is not considered. As shown in (7)-(10), in order to balance the thermal stress, the loss caused in each semiconductor must be adjusted. As in the case of the SM capacitor voltage, these losses depends on the SM switching state and arm current direction and magnitude. Considering the balance of the SM voltage, a complex control system or algorithm is required to achieve the optimal balance between SMs.

III. SM Capacitor Volage and Thermal Stress Balancing with RL based Modulation

A novel RL based NLM for MMCs is proposed for the optimal balance of junction temperature of semiconductors and SM DC capacitor voltage. As a first step to use RL, the system to be trained by RL should be defined by a Markov Decision Process (MDP) that is executed over a sequence of time-steps. MDP is based on agent-environment interactions through taking Actions (A_t) in the given State (S_t). The next state, S_{t+1}, is affected by the current state, S_t, and action, A_t, based on S_t. To satisfy MDP for RL, the state and action for 5 SMs in an upper arm of phase 'A' in 21-levels MMCs are defined in this paper as expressed in (11) and (12).

$$S_t = [N^*_{xu}, i_{arm}, V^N_{SM}, T^N_{JS1}, T^N_{JS2}, T^N_{JD1}, T^N_{JD2}] \tag{11}$$

$$a_t = [1, 2, 3, \cdots, 9] \tag{12}$$

The index of how well the actions are progressing is defined by the reward (R_t) function. The agent follows a Policy (strategy to pick actions) that maximizes its total accumulated reward over all time-steps. Therefore, choosing the correct agent and its learning condition for fulfilling appropriate modulation in MMCs for the multi-purposes is crucial. Based on the need for discrete actions, DQN is chosen here as a discrete-type agent with Q-value critic to take advantage of implementing the policy by a function approximator like neural network.

A. DQN Agent

The DQN agent uses a Q-function instead of using Q-table state-action pairs in a Q-learning algorithm because building a state-action-pairs table is computationally intractable, especially when the number of states is huge (e.g., here for 5 SMs is 27). Moreover, the DQN agent is suggested as it uses a loss function rather than an equation to map a state to the Q-values of all the actions that can be taken from that state, as shown in Fig.3. It has the advantage of improving predicted Q-value through comparing current Q-value to target Q-value. A deep neural network, as a nonlinear function approximator, is considered to predict action-value function Q(s, a). The network assigns a value to each possible action that can be taken in each state, which is input for the agent. The goal is to choose an action at certain state in order to maximize the Q-value, or the reward.

Like most of the RL algorithms, the basic idea of solving the proposed RL is finding optimal action-value function Q* (s, a) is based on Bellman equation. The basic form of this equation is represented in (13) where Q(s, a), Q(s', a') are state-action value for current state and next state respectively, γ is discount factor applied to weight later rewards over time for very long episodes to avoid infinitely growth of cumulative reward known as an expected total return (= $r + \gamma r_1 + \gamma^2 r_2 + \ldots + \gamma^n r_t$), and α is the learning rate in each episode.

$$\begin{aligned} &Q^N(\mathrm{s,a}) \\ &\leftarrow \mathrm{Q(s,a)} + \alpha\,[\mathrm{r(s,a)} + \gamma \max \mathrm{Q(s',a')} - \mathrm{Q(s,a)}] \end{aligned} \tag{13}$$

Using this equation allows to consider both immediate reward from the taken action to reach next state and cumulative discounted reward obtained from that state onward based on the policy. (14) represents how the optimal action-value function obeys the Bellman equation.

$$Q^*(s_t, a_t) = \mathbb{E}[r + \gamma \max(Q^*(s_{t+1}, a_{t+1}))] \quad (14)$$

Bellman equation estimates the action-value function in an iterative procedure, but separately for each sequence. To generalize the estimation, the mentioned neural network with weights θ is applied to estimate the action-value function, $f(s, a; \theta) \approx Q^*(s, a)$.

The loss function could be written as the difference between the target Q-value (y) and predicted Q-value ($\hat{y}$), (MSE (y, $\hat{y}$)). Assume that the prediction $\hat{y} = f(s, a; \theta)$, state-action value for next state $Q(s_{t+1}, a_{t+1})$ as $Q(s', a')$, and θ^- parameters of previous effort, then the loss function is obtained as

$$\begin{aligned} L(y, \hat{y}) &= L\big(Q^*(s, a), f(s, a; \Theta)\big) \\ &= \mathbb{E}\big[\big(r + \gamma \max\big(Q^*(s', a'; \theta^-)\big) - Q(s, a; \theta)\big)\big] \end{aligned} \quad (15)$$

The basic network is trained to change the weights, so that it minimizes a sequence of loss functions in each episode and leads converge to the optimal action-value function.

The Q-value critic is represented by a neural network having two-merged independent sequences layers of observations and actions. Four hidden layers with their activation layers are implemented to converge to the reward target. As the system needs consecutive approach, the sequence-layer is considered as an input neural network layer to take advantage of long short-term memory (LSTM) recurrent network, which makes the learning condition close to MDP. The last layer is a fully connected layer with a single neuron, representing the value.

B. *Reward function*

As described in above Section, the DQN agent in RL is based on the Q-function defined in (13) and acts to maximize the expected reward under the given state. In other words, the design of an appropriate reward function is one of the most important factors in reinforcement learning because the evaluation of state and action is performed with a reward function.

In order to train the agent, various reward functions have been applied in many studies. [10-14] Fig. 5 depicts commonly used four reward functions. As a basic function, there is the function providing constant and stepwise rewards when certain conditions are satisfied by actions of the agent. Although this method can be implemented intuitively, the problem of sparse reward may occur due to a discontinuous function. In contrast, it is also possible to apply a reward function based on continuous functions. Most basically, a reward function based on a first-order function can be applied. In this case, the reward can be linearly given to the agent according to the states. But, it is insufficient for encouraging the agent to reach optimal results. In order to compensate for this shortcoming, a nonlinear reward function, such as rational and inverse exponential functions, can be applied so that the agent to reach the optimal result effectively. In this paper, the inverse exponential function based reward functions are employed to balance the SM DC voltage and thermal stress.

Additionally, the error rate of the objective value to be balanced, SM DC voltage and semiconductor's temperature, rather than using the error value, as input to the exponential

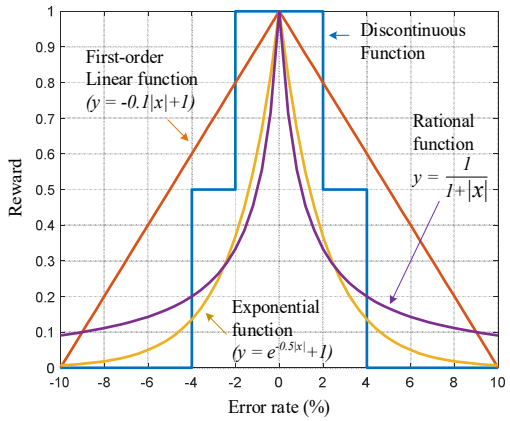


Fig. 5. Various reward functions

function. For example, if an error value is used as an input, the consistent reward cannot be given to the agent because the effect of the error is different depending on the average value of the target to be balanced. However, if the error rate is used, the certain reward can be given regardless of the value of the states. For example, the 40V error (10%) when the DC voltage is 400V and the 40V error (5%) when the voltage is 800V have different influences. In (15), a normalized equation about the objective values, X, that should be balanced is expressed.

$$E = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{X^i - avg(X)}{avg(X)} \times 100\right| \quad (16)$$

From (16), the reward function (r_t) for balancing the SM DC voltage (r_{vdc}), thermal stress ($r_{thermal}$) and common reward term are given in (17).

$$r_t = r_{vdc} + r_{thermal} + \frac{1}{2} r_{vdc} r_{thermal} \quad (17)$$

$$r_{vdc} = (2e^{-0.5E_{dc}} - 1) \quad (18)$$

$$r_{thermal} = \sum_{i=1}^{2} (e^{-0.5E_{JSi}} + e^{-0.5E_{JDi}}) \quad (19)$$

In case of the SM DC voltage, a penalty term is applied because the SM DC voltage is easy to be unbalanced and is not the preferred balance object. The third term is added for training the agent to encourage balancing two targets simultaneously.

IV. Simulation Results

Reinforcement Learning Toolbox in MATLAB is used to simulate the RL-based modulation to balance the SM DC capacitor voltage and thermal stress of semiconductors. Fig. 6 shows a block diagram configured in the simulation. As mentioned in the previous section, only five SMs in phase 'A' upper arm are operated by the agent to simplify the simulation model. Table I represents simulation and training conditions, respectively. In Table II, Epsilon, ε, is a parameter that the agent randomly selects an action during the training process so that the agent can explore the environment rather than output an action based on the Q-value function. If ε is 1, the agent outputs the action randomly with a 100% probability, and if it is 0, the agent acts according to the high Q-value according to the learned

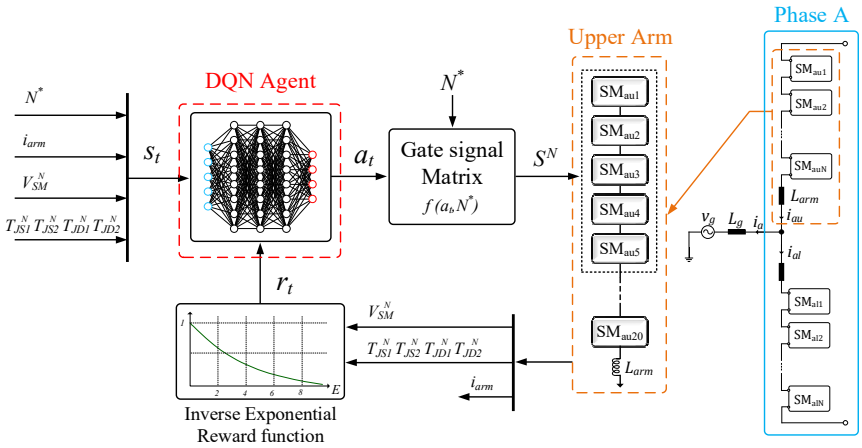


Fig. 6. Block diagram of proposed DQN agent based modulation for SM voltage and thermal stress balance in MMC.

TABLE I. SIMULATION AND TRAINING CONDITIONS

Symbol	Parameters	Value
-	SMs per Arm	20
C_{SM}	SM capacitance	15 mF
V_{DC}	DC voltage	8 kV
L_{arm}	Arm inductor	10 mH
L_{abc}	Phase inductor	5 mH
V_{grid}	Grid voltage	3.3 kV
α	Learning rate	$1e^{-4}$
γ	Discount factor	0.85
ε	Epsilon	1
ε_d	Epsilon decay	0.00005
ε_{min}	Epsilon minimum	0.01
-	Experience relay buffer	50000

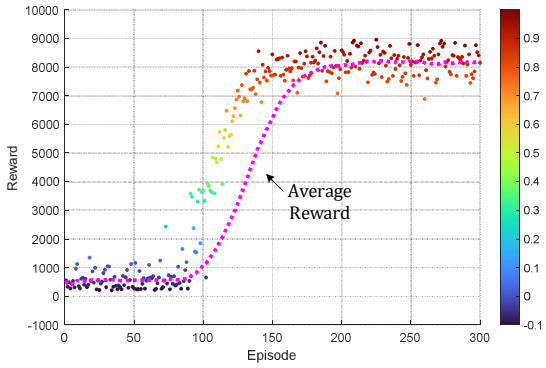


Fig. 7. RL agent training process

result. In Epsilon decay, ε_d, the value of ε decreases as the steps progress during the episode by ε_{min}. In other words, at the beginning of training, the agent is encouraged to explore the environment with random action. After the training has progressed, it is made to act according to the trained result. Experience relay buffer stores the training contents, and the larger the data buffer, the more precise training is possible by updating the neural network, but the training time becomes longer. These hyper-parameters should be fine-tuned as they affect the training time and precision.

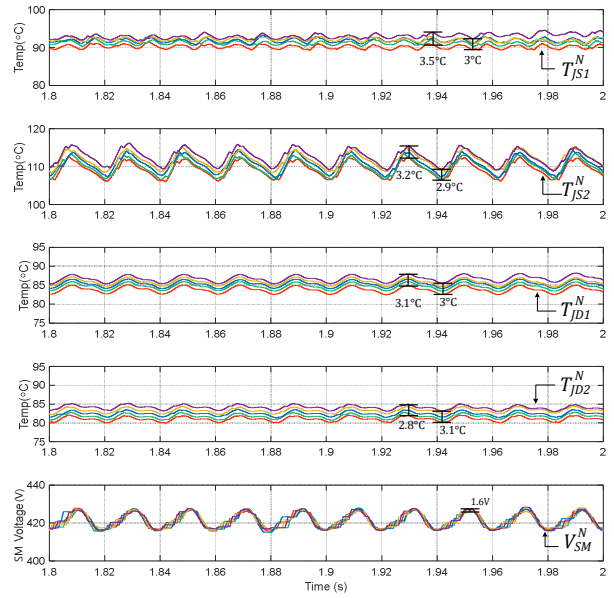
A single process of training the agent is conducted until the agent reaches to a target reward or certain maximum episodes. These training processes are repeated several times until the agent can achieve enough reward. This is because too many episodes in one training process is inefficient considering the epsilon decay and sometimes the reward falls suddenly to spoil the agent training. Consequently, a final training process of the agent is obtained as shown in Fig. 7. At the beginning of the training, the reward is very low because of the epsilon is close to 1. As the epsilon is decayed, the agent acts based on Q-value function which is a result of the training processes.

Fig. 8 shows the junction temperature of semiconductors and DC capacitor voltage of five SMs according to the modulation using the sorting algorithm to balance the SM voltage and the proposed modulation based on RL. In Fig. 8(a), errors between the temperature of the semiconductors are not balanced and their difference is about 3°C which is an initial error, even though the SM DC capacitor voltages are well balanced. On the other hand, when the proposed modulation is applied, as shown in Fig. 8(b), the junction temperature of diodes is balanced within 1°C and the temperature error of IGBTs is reduced by almost 1°C compared to the result in Fig. 8(a). In case of the SM voltage balance, it is balanced within 3% of the average voltage, which is acceptable difference, even though the balance of the SM voltages is worse than the result in Fig 8(a), because balancing the thermal stress is higher priority than balancing SM voltages and it can be adjusted by using the reward functions.

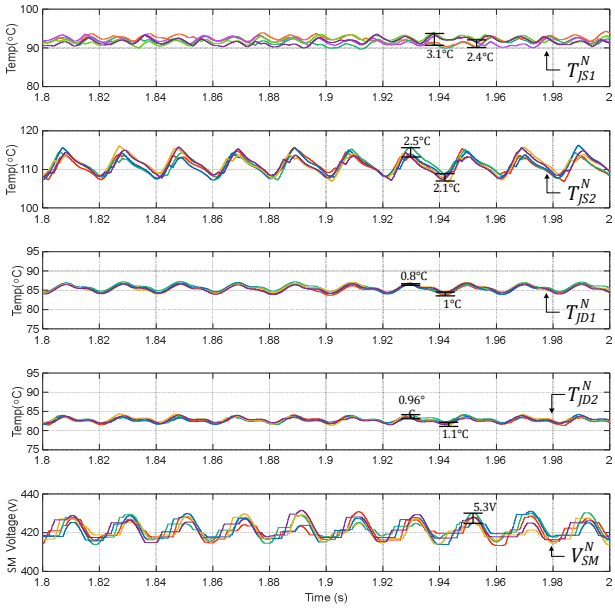
Fig. 9 compares episode rewards about the balance of junction temperature and SM voltage according to the modulations. As described in Fig. 8, it can be seen that the proposed modulation obtains a larger reward in terms of the temperature balance, but lacks the performance on the SM voltage balance. For a single episode (2000 steps), the average error rate between the junction temperature is less than 0.48% considering total accumulated rewards for the temperature balance is about 6300 and average reward for each step is 0.7875, $(e^{-0.5(0.4776)} = 0.7875)$.

V. CONCLUSION

Balancing the SM conditions is important for the NLM based MMC system which is composed of many SMs. Especially, the Imbalance of the thermal stress leads to the reduced current capability and different lifetime expectancy of semiconductors in MMC. To optimize the balance of the SM voltage and thermal stress, the DQN agent in RL is applied to NLM, and the reward function and hyper-parameters for training effectively the agent are presented. From simulation results, it was validated that the proposed RL based modulation can operate the MMC with more improved current capability by reducing the junction temperature errors between semiconductors in SMs than the general modulation, which cannot reduce the errors.

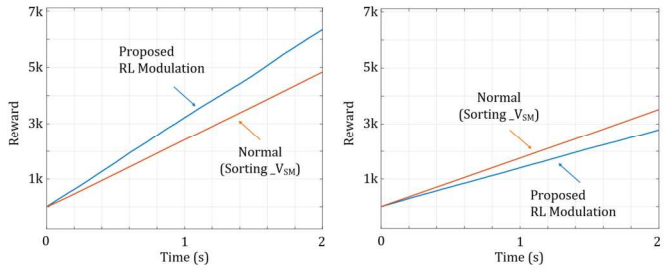


(a)



(b)

Fig. 8. Junction temperature and SM DC capacitor voltage results. (a) Normal modulation based on sorting algorithm for the SM voltage balance, (b) Proposed RL based modulation.



(a) (b)

Fig. 9. Accumulated reward comparison. (a) Junction temperature, (b) SM DC capacitor voltage.

ACKNOWLEDGMENT

The authors gratefully acknowledge funding by the German Federal Ministry of Education and Research (BMBF) within the Kopernikus Project ENSURE 'New ENergy grid StructURes for the German Energiewende' (03SFK1I0-2) and by the State of Schleswig-Holstein (22021016, KI-Förderrichtlinie) within the DatenCampus project.

REFERENCES

[1] S. Debnath, J. Qin, B. Bahrani, M. Saeedifard, and P. Barbosa, "Operation, control, and applications of the modular multilevel converter: A review," *IEEE Transactions on Power Electronics*, vol. 30, no. 1, pp. 37–53, 2015.

[2] G. Konstantinou, J. Pou, S. Ceballos, R. Darus, and V. G. Agelidis, "Switching frequency analysis of staircase-modulated modular multilevel converters and equivalent pwm techniques," *IEEE Transactions on Power Delivery*, vol. 31, no. 1, pp. 28–36, 2016.

[3] L. Lin, Y. Lin, Z. He, Y. Chen, J. Hu, and W. Li, "Improved nearest-level modulation for a modular multilevel converter with a lower submodule number," *IEEE Transactions on Power Electronics*, vol. 31, no. 8, pp. 5369–5377, 2016.

[4] P. M. Meshram and V. B. Borghate, "A simplified nearest level control (nlc) voltage balancing method for modular multilevel converter (mmc)," *IEEE Transactions on Power Electronics*, vol. 30, no. 1, pp. 450–462, 2015.

[5] M. H. Nguyen and S. Kwak, "Predictive nearest-level control algorithm for modular multilevel converters with reduced harmonic distortion," *IEEE Access*, vol. 9, pp. 4769–4783, 2021.

[6] J. Sheng, H. Yang, C. Li, M. Chen, W. Li, X. He, and X. Gu, "Active thermal control for hybrid modular multilevel converter under overmodulation operation," *IEEE Transactions on Power Electronics*, vol. 35, no. 4, pp. 4242–4255, 2020.

[7] F. Hahn, M. Andresen, and M. Liserre, "Enhanced current capability for modular multilevel converters by a combined sorting algorithm for capacitor voltages and semiconductor losses," in *2019 IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2019, pp. 3071–3077.

[8] F. Hahn, M. Andresen, G. Buticchi, and M. Liserre, "Thermal analysis and balancing for modular multilevel converters in hvdc applications," *IEEE Transactions on Power Electronics*, vol. 33, no. 3, pp. 1985–1996, 2018.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," arXiv, vol. abs/1312.5602, 2013.

[10] S. Zhao, F. Blaabjerg, and H. Wang, "An overview of artificial intelligence applications for power electronics," *IEEE Transactions on Power Electronics*, vol. 36, no. 4, pp. 4633–4658, 2021.

[11] Y. Tang, W. Hu, J. Xiao, Z. Chen, Q. Huang, Z. Chen, and F. Blaabjerg, "Reinforcement learning based efficiency optimization scheme for the dab dc–dc converter with triple-phase-shift modulation," *IEEE Transactions on Industrial Electronics,* vol. 68, no. 8, pp. 7350–7361, 2021.

[12] Ch. Cui, N. Yan, "An Intelligent Control Strategy for buck DC-DC Converter via Deep Reinforcement Learning", *Electrical Engineering and Systems Science*, Aug. 2020.

[13] P. Qashqai, K. Al-Haddad, R. Zgheib, "A New Model-Free Space Vector Modulation Technique for Multilevel Inverters Based On Deep Reinforcement Learning", *46th Annual Conference of the IEEE Industrial Electronics Society*, Singapore, Oct.2020.

[14] H. Jiang, Y. Chen, and Y. Kang, "Application of neural network controller and policy gradient reinforcement learning on modular multilevel converter - a proof of concept," in *2021 IEEE 4th International Electrical and Energy Conference (CIEEC)*, 2021.