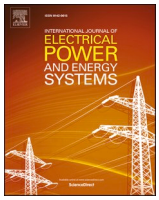
Electrical Power and Energy Systems 150 (2023) 109117

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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



Proximal policy optimization through a deep reinforcement learning framework for remedial action schemes of VSC-HVDC

Sungyoon Song [a], Yungun Jung [b], Gilsoo Jang [b], Seungmin Jung [c,*]

[a] Korea Electrotechnology Research institute, South Korea
[b] Korea University, South Korea
[c] Hanbat National University, South Korea

ARTICLE INFO

Keywords:
Artificial Intelligence
Proximal Policy Optimization
VSC-HVDC
Remedial Action Schemes
Energy Management System

ABSTRACT

A proximal policy optimization (PPO)-based back-to-back VSC-HVDC emergency control strategy based on multi-agent deep reinforcement learning (DRL) approach is proposed for use in an energy management system (EMS). In this scheme, an advanced DRL algorithm is proposed by implementing both PPO and a communication neural network for large power systems. The PPO modeled as intelligent agents with objective functions have shown a higher convergence performance than have existing DRL algorithms. Further, the model was demonstrated to effectively address voltage variances caused by the high penetration of renewable energy sources. By implementing PPO, the learning procedure is stabilized and made robust to continuous changes in network topology. To escalate the effectiveness of the proposed algorithm, a comprehensive case studies were conducted on an standard test systems and Korean power system considering variations in load and PV generation and a weak centralized communication environment. The results indicate that outstanding control performance and autonomously regulated bus voltage and line flows, thereby validating the effectiveness of the method.

1. Introduction

The contribution of renewable energy sources (RESs) to electrical networks has grown considerably in recent years, causing power system operators concern about the impact of RESs on the operational security and efficiency of their networks [1]. The high penetration of RESs has led to new challenges in planning and operating transmission systems, including system operation cost, equipment overloads, and voltage quality problems. In South Korea, most RESs are not yet connected to the grid. Although the Korean government has mandated 77 GW of RESs to be installed by 2034 [2] and much progress has been made toward developing RESs, existing transmission systems cannot effectively address the voltage regulation and operation cost issues caused by the high penetration of intermittent RESs owing to their limited regulation and project delays. System operator is trying to solve the voltage issues by the unified module which is combined with the stability prediction tool and the unit-commitment (UC) algorithms. However, the voltage problems were not mitigated by the UC algorithms because generators are located further away from the metropolitan areas. Alternatively, optimal power flow (OPF) module which can regulate switched shunts and transformers has been implemented for voltage regulation. However, gradually increasing the prediction error of RESs will lead to accelerated inaccurate control output. In 2030, ±0.6 GW variation can occur when there is a 5% of forecast error in South Korea.

One solution that has been proposed to address these issues is the use of closed-loop systems with back-to-back (BTB) voltage source converter (VSC) based high voltage direct current (HVDC) in the city. A single ac network fault can be propagated quickly over a wide area [3]. Alternatively, BTB VSC-HVDC is power electronic device that can replace normally open/closed points to provide a fast response, frequent actions, and an enhanced control scheme for bidirectional power flow between adjacent substations [4]. Prior researchers have demonstrated the capability of VSC-HVDC to increase the penetration of RESs without upgrading existing transmission lines via multiple control functions [5–8].

The transmission technology based on BTB VSC-HVDC has many advantages as follows: realizes the independent control of active and reactive power; provides black start capability and dynamic reactive power support [9]; regulates the network voltage profile [10]; divides the ac system; and is suitable for grid-connected weak and urban grids [11]. Several researchers have also focused on designing an optimal control strategy, e.g., by minimizing energy losses [12] or annual system expenses, enhancing the voltage profile [13]. Recently, decentralized

* Corresponding author.
E-mail address: seungminj@hanbat.ac.kr (S. Jung).

https://doi.org/10.1016/j.ijepes.2023.109117
Received 26 September 2022; Received in revised form 10 February 2023; Accepted 27 March 2023
Available online 4 April 2023

Nomenclature

v_1	Voltage at the converter
v_2	Voltage at the point common coupling
R	Resistance
C, L	LC filter capacitance and inductance
i_1	Currents flowing through the ac system
i_2	Three-phase current flowing through the inductor
ω	Angular frequency of the ac voltage
Δe_p	Errors of the active power controller
Δe_v	Errors of voltage controller
k_{PLL}	Phase lock loop gain
A	State matrix of VSC
J	Total generation costs
k	Total number of generators
$C_{i,h}$	Generation cost curve at hour h
$P_{i,h}$	Active power generated at generator bus i and hour h
P_{injn}	Total active power at bus n and hour h
Q_{injn}	Total reactive power at bus n and hour h
V_n	Magnitudes of the voltages at buses n
V_l	Magnitudes of the voltages at buses l
G_{nl}	Conductance of the admittance
B_{nl}	Susceptance of the admittance
δ_l	Voltage angles at buses n
δ_n	Voltage angles at buses l
P_{linel}	Power flow limit of the line connecting bus n and l
P_{subl}	Power flow limit of the substation
S, A, O	States and action and observations
r	Rewards
T	Transition function
p	Transition model
π, θ	Policy and policy parameter
π_θ^*	Optimal policy
d	Discount factor
E_π	Expectation
G_t	Total expected return
r	Probability ratio
p_x	Active power flow at the VSC-HVDC
q_x	Reactive power flow at the inverter
z_x	Reactive power flow at the rectifier
S	Converter rated capacity
D	Demand data
K	Renewable energy data
$\mathscr{L}$	set of states
v_t^{rec}, v_t^{inv}	Measured voltage at target feeders
α, β, γ	User-defined constants for reward shaping

control strategies have been addressed as follows: master–slave control strategy [14], voltage margin control [15], dc voltage droop control [16] and adaptive droop control [17,18]. However, accuracy can be compromised, particularly when new RESs are connected to the VSC because the current standard is to employ proportional control loops locally at converter. The timetable or three-dimensional fitting curves should also be updated to prepare for abrupt power changes of RESs.

In this study, therefore, deep reinforcement learning (DRL) framework for VSC-HVDC with the widely used remedial action schemes (RAS) is introduced. Successful operation of two converters (rectifier and inverter) requires flexibility in the managing of rapid and large voltage deviations arising from sudden generations of RESs. However, existing rule-based RAS have employed representative values for loads, renewable generations, and network topology, leading to inaccurate results in real power systems. Accuracy also suffers when rapid topology changes occur in the network during optimization using strategies to solve convex or non-convex optimization problems [19]. For example, UC algorithms uses a mixed-integer nonlinear optimization method which takes longer times to solve. Thus, real-time optimal values cannot be derived because the high penetration of intermittent RESs accelerates power system topology changes.

Furthermore, current energy management system (EMS) and situational awareness system (SAS) have used instantaneous measurements with the Markov property as inputs to the control algorithms. However, the dynamic response of a power system cannot be modeled as a first-order Markov model [20]. The future state of a system depends on a range of unknown state variables. As existing methods have relied heavily on *a priori* knowledge of complex and large power systems, many model-based methods have difficulties handling uncertainties from stochastic changes inherent to some RESs [21]. The centralized voltage control, which is based on an OPF, is performed with a 5-minute time interval. Nevertheless, if the voltage deviations arising from the sudden generations of RESs, the proposed RAS is activated using DRL model for real-time control. Inaccurate RESs prediction force the bus voltages deviate from their nominal values. If the severe voltage stability issues arise, the RAS is activated until the next OPF cycle.

Some researchers have already attempted to incorporate DRL algorithms to improve the performance in power system operation [22,23,24,25]. With interest to develop real-time recommendation systems, artificial intelligence based controllers have been of interest to the industry. A DRL framework with volt-var optimization algorithms have been proposed [26] and a hierarchical deep deterministic policy gradient algorithm for automatic generation control has been proposed [27]. However, all these references mostly discuss how DRL can be implemented to power system operation.

Recently, model-free DRL has been addressed [28]; but, it is impractical for use with function approximation and suffers from training stability issue. Moreover, the actor-critic [29] has been widely addressed because the size of target network is small. However, there is an increasing demand for transmission system operators to be able to control multiple devices in real-time, not only single devices. There is a lot of state variables in power system environment, which are essential for DRL process. Increased state variables can degrade training stability and for this reason are considered as a potential risk factor. Simply stated, the full observation vector of the power system has detailed information about the status of the grid, but the large number of the observations increases training instability. The training stability can be compromised if the algorithm cannot prevent new policy from straying too far from the old policy; thus, a stable learning algorithm should be implemented in real system operation.

Considering these factors, this study proposes a proximal policy optimization (PPO)-based DRL architecture, which can achieve robust voltage control for VSC-HVDC. Note that we refer the autonomous control framework "*Grid Mind* [24,25]" which controls agent trained using DRL algorithms by interacting with numerous calculations of a power flow. Multi-agent DRL (MADRL) algorithms with PPO modeled as intelligent agents with objective functions have shown a higher convergence performance than have existing algorithms. By employing the existing EMS configuration, the PPO-based RAS model can be added, as shown in Fig. 1.

Overall, the proposed model should thus be straightforward to implement, tolerant to additional coordination settings. Therefore, it is an attractive interim option before implementing more complex optimization algorithms. In summary, four main contributions are targeted, listed below.

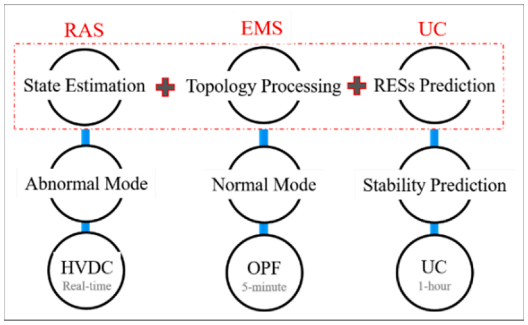


Fig. 1. EMS, UC, and proposed RAS model for RESs operation.

1) VSC-HVDC is firstly used in a goal-oriented control scheme with a DRL model, thereby achieve better total voltage regulation
2) To configure the reward functions, an iterative two parallel OPF calculation are implemented in the learning framework.
3) To improve training stability, PPO is used for learning algorithms, which adopts the trust-region concept.
4) Advanced RAS architecture allows flexibility and redundancy even with unforeseen RESs power change.

These efforts are detailed in the following sections. In Section 2, the control stability of VSC-HVDC is introduced. In Section 3, OPF and reinforcement learning algorithms are described. The PPO-based DRL with training platform is illustrated in Section 4. Finally, a simulation of the developed model with the proposed control scheme is presented in Section 5, and conclusion is presented in Section 6.

2. Control stability of VSC-HVDC

Several researchers have noted that a VSC-HVDC with large gains that is connected to a weak ac grid is prone to instabilities when subjected to a disturbance [30]. The ac system connected to the converter is modeled in a synchronously rotating reference *d-q* frame, and the *q*-axis is locked with the ac voltage to ensure decoupled control. The dynamics of the ac side of the converter in the *d-q* frame can be expressed as follows:

$$\begin{bmatrix} v_1^d \\ v_1^q \end{bmatrix} - \begin{bmatrix} v_2^d \\ v_2^q \end{bmatrix} = R\begin{bmatrix} i_1^d \\ i_1^q \end{bmatrix} + L\frac{d}{dt}\begin{bmatrix} i_1^d \\ i_1^q \end{bmatrix} + \begin{bmatrix} -\omega L i_1^q \\ \omega L i_1^d \end{bmatrix}, \tag{1}$$

$$\begin{bmatrix} i_1^d \\ i_1^q \end{bmatrix} - \begin{bmatrix} i_2^d \\ i_2^q \end{bmatrix} = C\frac{d}{dt}\begin{bmatrix} v_2^d \\ v_2^q \end{bmatrix} + \begin{bmatrix} -\omega C v_2^q \\ \omega C v_2^d \end{bmatrix}. \tag{2}$$

where v_2 is the three-phase output voltage at the point common coupling (PCC) and v_1 is the voltage at the converter. Furthermore, R, C and L are the resistance, LC filter capacitance, and inductance, respectively; while i_1 and i_2 are the currents flowing through the ac system and the three-phase current flowing through the inductor, respectively. The symbol ω is the angular frequency of the ac voltage at the PCC. Based on (1) and (2), the reduced third-order small-signal model of a single VSC can be written as follows:

$$\frac{d}{dt}\begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\delta \end{bmatrix} = \mathrm{A}\begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta\delta \end{bmatrix}. \tag{3}$$

$$\frac{d\Delta x_1}{dt} = \Delta e_p, \ \frac{d\Delta x_2}{dt} = \Delta e_v, \ \frac{d\Delta\delta}{dt} = k_{PLL}\Delta v_2^q. \tag{4}$$

where Δe_p and Δe_v are the errors of the active power controller and voltage controller, respectively; and k_{PLL} is the PLL gain. The coefficient parameters of the matrix A can be found in Appendix [31]. Based on three linearized state variables and the state matrix components, it was confirmed that ac voltage controller is strongly coupled with the PLL. To analysis of the mutual interaction between the ac voltage controller and the PLL, the roots of the quadratic equation are calculated. As shown in Fig. 2, the results show that real-axis eigenvalue as each parameter was increased 10-fold according to the two grid strengths.

where k_i^v and k_p^v are the PI parameters of the ac voltage controller; k_p^P is the gain of the power controller. Under the precondition that all the real parts of the characteristic roots were laid on a negative value to avoid the underdamped response, the sensitive parameters should be able to shift the root positions. The parameter with a large negative y-axis value can easily shift the system to the left half-plane. As shown in Fig. 2, the parameter that contributes the most to the system stability is the time constant of voltage controller; thus, the ac voltage phase is highly sensitive to the *q*-axis current injections of the converter according to the system operating points. For example, the low-pass filter structure is used to solve the small signal stability problem caused by the fast response of the inner current loop in grid-following converter. In conclustion, the frequent change of voltage or reactive power reference can impact the converter stability in weak grid; thus, the proposed control mode is to act as an emergency control with RAS that should be activated in abnormal mode.

3. Training environment

The proposed real-time control method is based on the offline training of a PPO-based DRL algorithm on a large dataset consisting of an OPF following a set of generation costs. The proposed architecture allows multiple agents to be trained offline by interacting with massive offline simulations and historical events. The overall system architecture is illustrated in Fig. 3.

A. *OPF Model in Training Environment.*

Firstly, the probability model of the RESs is configured to calculate OPF. In OPF calculation, a probability model of RESs is implemented to evaluate the influence of the uncertainty of RESs outputs and load on the running state of the power system. Because irradiance is uncontrollable and stochastic, injections of power into the power system become more stochastic, causing stochastic variations of the system state. The probability distributions of the output stochastic terms are obtained by their probability density function [32]. The object function is formulated by minimizing the total generation cost as follows:

$$MinimizeJ = \sum_{h=1}^{8760}\sum_{i=1}^{k} C_{i,h}\left(P_{i,h}\right), \tag{5}$$

$$C_{i,h}\left(P_{i,h}\right) = a_i \times P_{i,h}^2 + b_i \times P_{i,h}^2 + c_i. \tag{6}$$

where J is the total generation costs in the power system, k is the total number of generators, and $P_{i,h}$ is the active power generated at generator bus i and hour h; the latter term is stochastic owing to the presence of intermittent RESs. The cost function is represented as a quadratic function of the generator output active power, as shown in (6). And, the

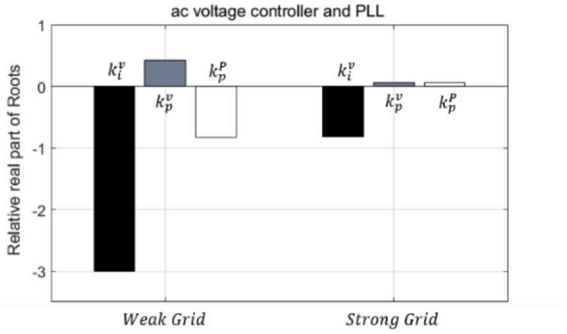


Fig. 2. Result of the sensitivity analysis of ac-voltage control and PLL.

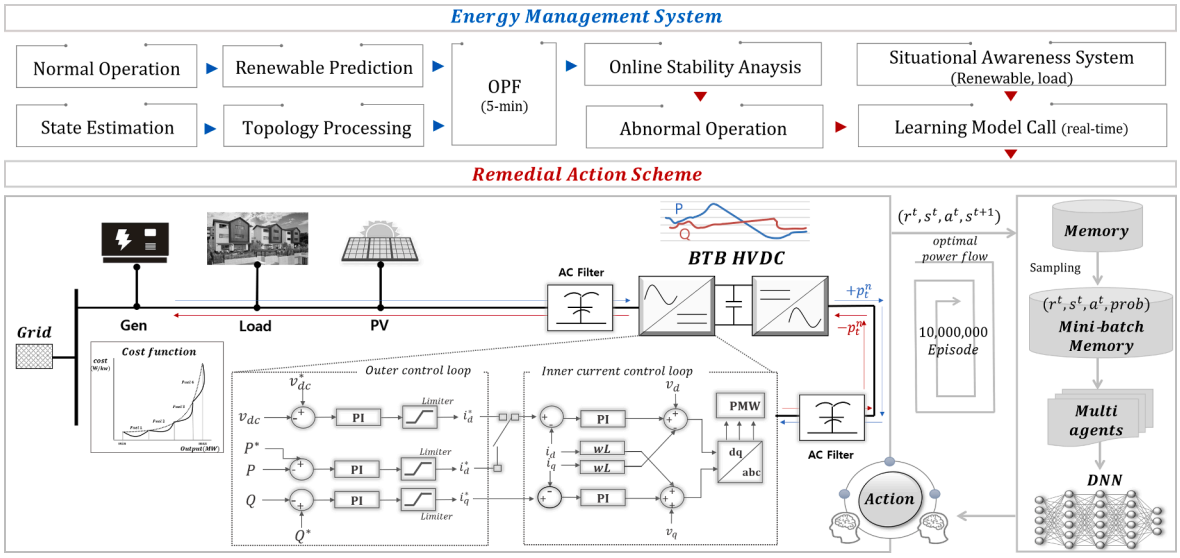


Fig. 3. System architecture.

OPF was constrained according to:

$$P_{injn,m} - \sum_{l=1}^{N} V_{n,m} \times V_{l,m} \times \left[G_{nl} \times cos\left(\delta_{l,m} - \delta_{n,m}\right) + B_{nl} \times sin\left(\delta_{l,m} - \delta_{n,m}\right)\right] = 0, \tag{7}$$

$$Q_{injn,m} - \sum_{l=1}^{N} V_{n,m} \times V_{l,m} \times \left[G_{nl} \times sin\left(\delta_{l,m} - \delta_{n,m}\right) + B_{nl} \times cos\left(\delta_{l,m} - \delta_{n,m}\right)\right] = 0, \tag{8}$$

where $P_{injn,m}$ and $Q_{injn,m}$ represent the total active and reactive power injected into the power system at bus n and hour h, respectively, $V_{n,m}$ and $V_{l,m}$ are the magnitudes of the voltages at buses n and l, respectively, G_{nl} and B_{nl} are the conductance and susceptance of the admittance, respectively, and $\delta_{l,m}$ and $\delta_{n,m}$ are the voltage angles at buses n and l, respectively.

$$V_{min} \le V_i \le V_{max}, i = 1, 2, \cdots, N \tag{9}$$

$$P_{min} \le P_i \le P_{max}, i = 1, 2, \cdots, N \tag{10}$$

$$Q_{min} \le Q_i \le Q_{max}, i = 1, 2, \cdots, N \tag{11}$$

(9)–(11) show the voltage magnitudes and their limits at buses, active power and reactive power output limits at the generator and branch flow limit. The power flow was constrained as:

$$\sum_{l=1}^{N} V_{n,m} \times V_{l,m} \times \left[G_{nl} \times cos\left(\delta_{l,m} - \delta_{n,m}\right) + B_{nl} \times \sin\left(\delta_{l,m} - \delta_{n,m}\right)\right] \le P_{subl}, n, l = 1, 2\cdots, N, \tag{12}$$

$$\sum_{l=1}^{N} V_{n,m} \times V_{l,m} \times \left[G_{nl} \times cos\left(\delta_{l,m} - \delta_{n,m}\right) + B_{nl} \times \sin\left(\delta_{l,m} - \delta_{n,m}\right)\right] \le P_{linen,l}, n, l = 1, 2\cdots, N. \tag{13}$$

where P_{linel} is the power flow limit of the line connecting bus n and l, and P_{subl} is the power flow limit of the substation. The OPF performs economic dispatch at every time step by considering several constraint equations, and the probability behavior of RESs is simultaneously considered in the learning process.

B. Preliminaries of Markov Games [33].

A multi-agent extension of Markov decision processes (MDP) can be described by Markov games. An MDP is defined by a set of states $s^t \in S$, actions $a_i^t \in A_i$, and observations $o_i^t \in O_i$ for each agent. A transition function, $T : S \times A \to P(S)$, defines the effects of the probability actions on the state of the environment. When each agent performs an action, the environment changes as a result of the joint action according to the state transition model $p(s^{t+1}|s^t, a^t)$. Each agent obtains rewards as a function of the state and the joint action: $r_i^t : S \times A \to \mathbb{R}$, with $r^t = r^t(s^t, a^t, s^{t+1})$. Solving an MDP thus involves finding the individual policy $\pi_i : O_i \times A_i \to [0, 1]$. Each agent receives an observation o_i^{t+1}. The π is parameterized by θ, which is determined by the weights and biases of a neural network. The goal is to find a policy that maximizes the cumulative discounted rewards. When the T-step trajectory is defined as $J(\pi_i) = \sum_{t=0}^{T} d^t r_i^t$, where $d^t \in [0, 1]$ is a discount factor, the objective can be expressed as:

$$\pi_\theta^* = \underset{\theta}{\text{argmax}} J(\pi_\theta). \tag{14}$$

Here, π_θ^* represents the optimal policy. Once all agents complete their actions, the algorithm transfers to the next state.

C. *Preliminaries of Proximal Policy Optimization.*

PPO is based on the Monte Carlo policy gradient theorem [33], which is used to find the optimal reward from each agent in behavior strategy. By using estimated return, the policy parameter θ is updated iteratively using the Monte Carlo method. As the total expected return G_t can be calculated from the real sample trajectory, the policy gradient is updated as:

$$\nabla_\theta J(\theta) = \mathrm{E}_\pi[Q^\pi(s, a) \nabla_\theta In\pi_\theta(a|s)] = \mathrm{E}_\pi[G_t \nabla_\theta In\pi_\theta(A^t|S^t)]. \tag{15}$$

The Monte Carlo method uses the entire sample trajectory to update the policy gradient. The intuitive procedure follows four steps: 1) the policy parameter θ is randomly selected, 2) the trajectory is formulated using policy π_θ, 3) the total expected return G_t is estimated at $t = 1, 2, \cdots T$, and 4) the policy parameter is updated as $\theta \leftarrow \theta + ar^t G_t \nabla_\theta In\pi_\theta(A^t|S^t)$.

The PPO has thus been derived as a soft-constraint version of Kullback–Leibler divergence, combines policy with a safety intervention module, and adopts a trust region to improve training stability by ensuring a clipped surrogate objective at every iteration. In this algorithm, the probability ratio of the old and new policies is:

$$r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}. \tag{16}$$

The on-policy objective function can then be represented as:

$$J^{PPO}(\theta) = \mathrm{E}\left[r(\theta) \widehat{A}_{\theta_{old}}(s, a)\right]. \tag{17}$$

If there are no upper and lower bounds between θ_{old} and θ, the policy update to maximize $J^{PPO}(\theta)$ can be very large; large update steps taken along the trajectory can cause the algorithm to become trapped in the circled valley of the objective function geometry. To avoid this, $r(\theta)$ is constrained within $[1-\epsilon, 1+\epsilon]$. The hyperparameter ϵ prevents the new policy from straying too far from the old policy. The clipped objective function can thus be represented as:

$$J^{CLIP}(\theta) = \mathrm{E}\left[\min\left(r(\theta)\widehat{A}_{\theta_{old}}(s,a), clip(r(\theta), 1-\epsilon, 1+\epsilon)\widehat{A}_{\theta_{old}}(s,a)\right)\right]. \tag{18}$$

where ϵ is used to regulate the size of the trust region in the update. If $\widehat{A}_{\theta_{old}}$ is positive and $\pi_\theta(a|s)$ increases, then $J^{CLIP}(\theta)$ increases, the old policy is favored, and the minimum operation determines the ceiling, i.e., $(1+\epsilon)\times\widehat{A}_{\theta_{old}}$. During PPO, the lesser value between the original and clipped values is selected. Thus, there is no action to make the policy extremely small for better rewards.

4. PPO model for Real-time VSC-HVDC control

In the proposed training platform, multi-agents were trained offline via numerous simulations to prevent unexpected power generation changes of RESs. Each agent had its individual actor, critic, coordinator, and replay buffer, as shown in Fig. 4. Agents could share the state-space information via a CommNet [34], which uses continuous communication for fully cooperative tasks and the convergence issue has been demonstrated to perform significantly better than does an independent controller. By passing an averaged message over agent modules between layers, a single network can be used in a multi-agent setting. The CommNet is used to obtain an integrated communication vector for each agent by average pooling the messages broadcast from the agents. And, the offline agent training process for designing the PPO-based model, including state, action, and reward, is discussed in this section.

A. *State.*

The state space for training the agent consists of the state of two converters, total generation cost, and load/RESs profiles. The power flow direction set in the VSC-HVDC at time step t is denoted as $\boldsymbol{p}_x = \{\pm p_t^0, \pm p_t^1, \cdots, \pm p_t^k, \cdots \pm p_t^{K-1}\}$, where p_t^k represents the active power flow control results at the k th VSC-HVDCs and $\pm$ denotes the direction of power flow. The converters perform constant reactive power control as $\boldsymbol{q}_x = \{q_t^0, q_t^1, \cdots, q_t^k, \cdots q_t^{K-1}\}$ and $z_x = \{z_t^0, z_t^1, \cdots, z_t^k, \cdots z_t^{K-1}\}$ at the rectifier and inverter, respectively. The converter capacity was constrained as $S_{inv} = \sqrt{p_t{}^2 + q_t{}^2}$ or $S_{rec} = \sqrt{p_t{}^2 + z_t{}^2}$, where q_t and z_t represent the reactive power for voltage regulation at time step t and S is the converter rated capacity.

The total system loss and generation cost is changed when the converter control, and the set of demands and RESs at time step t were defined as $\boldsymbol{D} = \left\{d_t^0, d_t^1, \cdots, d_t^a, \cdots d_t^{A-1}\right\}, \mathbf{K} = \left\{k_t^0, k_t^1, \cdots, k_t^b, \cdots k_t^{B-1}\right\}$, where d_t^a and k_t^b represent the load at a th bus and the RESs generation at b th bus, respectively. In real power system, the data $\boldsymbol{D}$ and $\mathbf{K}$ presented above are measured as the SAS. Furthermore, the voltage measured by the phasor-measurement-unit is denoted as $\boldsymbol{v}_x = \left\{v_t^0, v_t^1, \cdots, v_t^k, \cdots v_t^{C-1}\right\}$, where v_t^k represents the voltage at c th bus. The voltage profile is smoothed by the low-pass-filter as described in section D. The conventional generators also have their respective cost functions and their power is regulated to satisfy net load at time step t. The set of states at time step t is denoted as $\mathscr{L}_\mathbf{t} = \left\{s_t^0, s_t^1, \cdots, s_t^D, \cdots s_t^{D-1}\right\}$, where $s_t^D = \left[d_t^a, k_t^b, v_t^c, cost_t\right]$. Here, $cost_t$ is the total generation cost at time t, which is determined by the interior point method for re-dispatching generators in the OPF model. The cost state is calculated as $\sum_{i=1}^{24} cost$ in each episode and chosen as a reward.

B. *Action.*

To satisfy the net demand at each time step t, the generator at the slack bus automatically regulates the power amount at the feeder point to balance any power imbalances in the power network. Thus, a high generation cost was assumed in the main grid in the OPF model. Each generator regulates its active power to minimize total system cost. And, two converters strive to change both power flow direction and voltage, each converter were thus selected as the agent's action space $\left\{\pm p_t^0, \pm p_t^1, \cdots, \pm p_t^k, \cdots \pm p_t^{K-1}\right\}$, $\left\{q_t^0, q_t^1, \cdots, q_t^k, \cdots q_t^{K-1}\right\}$, $\left\{z_t^0, z_t^1, \cdots, z_t^k, \cdots z_t^{K-1}\right\}$.

C. *Reward.*

To evaluate the effectiveness of the actions of two agents, a reward function comprising three objectives was implemented in (19)–(21). First, the voltage deviation was minimized via r_{y1}^t and r_{y2}^t, which is determined by the reactive power control of two converters (for action $\boldsymbol{q}_x$ and z_x). Note that the reason for applying reactive power control rather than voltage control is to fully utilize active power range under converter capacity constraint. The converter takes priority of active power control and uses the rest for reactive power control. Third, the cost benefits provided by the active power control of VSC-HVDC were maximized via r_z^t, as the system cost varies with agent action, especially under the power generations caused by RESs. Moreover, the hidden reward function which has a large negative value has been implemented when the power flow is not converged or divergence. To ensure the robustness of agent, a threshold of 0.99 or 1.01 is used instead of 1.0 during training to the agent. The reward functions impact both the training speed and convergence; therefore, the rewards multiplied by suitable coefficients were described by (19)–(21).

$$r_{y1}^t = \beta \times -\left(\sqrt{\left|1.0 - v_t^{rec}\right|^2} - \sqrt{\left|1.0 - v_{t,noact}^{rec}\right|^2}\right), \tag{19}$$

$$r_{y2}^t = \beta \times -\left(\sqrt{\left|1.0 - v_t^{inv}\right|^2} - \sqrt{\left|1.0 - v_{t,noact}^{inv}\right|^2}\right), \tag{20}$$

$$r_z^t = \gamma \times -\left(cost_t - cost_{t,noact}\right). \tag{21}$$

$$r^t = r_{y1}^t + r_{y2}^t + r_z^t \tag{22}$$

Here, v_t^{rec} and v_t^{inv} are the measured voltage at target feeders, which means that the number of v_t^{rec} and v_t^{inv} depends on the system size. For example, the number of v_t^{rec} was set to 5 and 6 for v_t^{inv} was applied in the IEEE-39 test system; thus, a total of 10 voltage rewards is considered. $v_{t,noact}$ and $cost_{t,noact}$ represent the voltage and total system cost when the agents are not performing at time step t, and are calculated by an internal parallel OPF loop in the proposed learning framework to enhance the reward shaping. α, β, γ are user-defined constants representing the weight of each of the three reward terms.

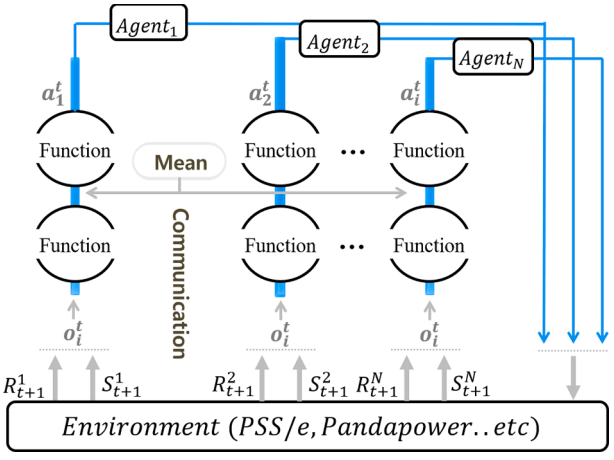


Fig. 4. Training architecture for the proposed PPO algorithm.

The proposed training algorithm, summarized in Table 1, is composed of three parts, in which the probability function of each RES, the OPF, and the drive status and reward are calculated. The closed-loop training dataset comprised 1,000 episodes, where each episode comprised 5-min intervals (for a total of 24 hrs) of raw data; for each episode, the corresponding decisions and rewards were saved.

D. *Proposed Control Scheme of VSC-HVDC.*

As mentioned in section A, many voltage fluctuations are recorded by the phasor-measurement-unit, which has a high sampling rate (50 samples/sec). The voltage data is used for the input data in real RAS operation. The data must be smoothed before injecting into the PPO model because the discrete values were implemented in the learning process. Smoothing of the voltage profile through a low-pass-filter can be achieved by $1/\ 1 + T_s s$. Increased the value of T_s results in smoother output voltage and also longer the time delay; thus, the larger T_s was used in this algorithm. An infinite value indicates that inaccurate data was measured; thus, invalid data was replaced by the one-step old value.

The final control schemes of VSC-HVDC is illustrated in Fig. 5. In normal mode, the control output values are calculated by OPF in EMS, and the proposed DRL architecture is activated in abnormal mode when there is a large voltage deviation. Gradually increasing the forecast error of RESs will lead to accelerated inaccurate control output; thus, the proposed DRL model can overcome this challenge because the model uses *current* generation amount of RESs.

5. Results and Discussion

A. *Simulation Setup.*

To examine the performance of the proposed method, simulations were performed in the IEEE-39 bus test system comprising 39 buses and 28 branches, as shown in Fig. 6. The PV systems including their respective stochastic model were attached at buses 4, 16, 17, 18, 19 and 21. The PV power data supplied was collected by the Korea Institute of Energy Research from year. The stochastic PV data and the load were properly scaled to apply the IEEE-39 bus standard system, as shown in Fig. 7. The PVs have generation capacity of 80 MVA and the converter has capacity of 200 MVA. The piecewise linear cost functions [35] of the generators used in the IEEE-39 bus system are given in Table 2.

The program codes were written and compiled in the Python 3.7 environment using TensorFlow 1.13.0 to configure the algorithm. Additionally, pandapower 2.3.0 elements and functions were used to implement the OPF in the IEEE-39 bus standard test system. All simulations were conducted using a personal computer equipped with a 3.50-GHz Intel® CoreTM i9-11900 K central processing unit, 32 GB of random-access memory, and a 64-bit Windows® 10 operating system.

By considering IEEE-39 bus system size, we choose 10 kinds of respective action from a discrete finite set for each agent. For example, the discrete action of the two converters, the active power and reactive power for both side were set to $\boldsymbol{p} = \{-50, -90, \cdots, 90, 50\}$, $\boldsymbol{q} = \{-100, -90, \cdots, 90, 100\}$, $z = \{-100, -90, \cdots, 90, 100\}$. The simulation was executed for 24 h with a 5-minute time interval.

Table 1
PPO-based Reinforcement Learning Algorithm.

Algorithm PPO-based MADRL Training Algorithm
1: Initialize optimal power flow and target network
2: Initialize the critic and actor networks with weights
3: **for** episode = 1 to N **do**
4: **for** time step = 0 h to 24 h **do**
5: Calculate actions a_i^t for each $s^t \in S$
6: Execute a_i^t in OPF solver environment
7: Execute OPF solver environment without a_i^t
8: Send a set of states s^{t+1} and reward r^t to each agent
9: Select a_i^{t+1} through a_i^t, s^{t+1}, r^t
10: Store the transition pairs in replay buffer
11: Sample a random minibatch
12: Update target critic and actor by stochastic gradient descent to the loss function of network
13: Step+ =1
14: **end**
15: **end**

Regarding the parameters of PPO model, the Adam optimizer has been implemented. There were three layers with 256 hidden layers, as shown in Table 3. The clip which regulates the size of the trust region is determined by 0.1 for training stability. Several researchers have already noted that tuning the reward function is a challenge for deep reinforcement learning model. It requires a combination of heuristics based on prior knowledge and some automated parameter search. A total of 20 unique training scenarios are selected from EMS data, and the agent continuously regulates voltage for all time steps in a scenario is categorized as a successful agent. The following results are used to verify hyper-parameters.

B. *Performance Evaluation.*

To evaluate the effectiveness of the proposed model, comparative tests were carried out on test condition. Two control schemes were compared, including the actor-critic [35] and PPO. The resulting reward convergence tendency of the training curve under the two DRL algorithms is shown in Fig. 8. The PPO algorithm indicated by the blue line is rapidly converged with small variances. It shows significantly better than does the use of other DRL algorithms on a multi-agent system. Apparently, training stability can be achieved at any time and meaningful learning model can be made even with a large power system environment. This has an appealing advantage of having the ability to operate stably in EMS. The actor-critic method also converged well; however, the algorithm cannot prevent the new policy from straying too far from the old policy sometimes so that the reward variation is large. As the power system environment becomes more complex, the stability of algorithm with the actor-critic method falls below that of the proposed algorithm. Overall, the proposed algorithm demonstrated a high reward convergence and thereby derived an optimized policy. Also, in real power system, the total bus splitting/merging topologies at a substation with k elements is 2^{k-1}. Thus, the problem is there is so many possible topology configurations available at a given time step. PPO deals with VSC-HVDC controllers for sequential decision processes to achieve specific objective. Therefore, the total number of topology configurations has been reduced to validate control actions. Representative hourly topology configurations were chosen from the EMS data. The observations from representative topology configuration are used as inputs by the RL agent to determine the action in the next time step.

C. *Simulation Results.*

1) CASE 1 - The actions for two agents and PV profiles obtained for the IEEE 39-bus test system indicate that the action profile of the proposed method is effective with high reward convergence result. By observing the reactive power control actions, as shown by pink and blue lines in Fig. 9(b), the proposed algorithm fully utilized its control range and intentionally regulates its bus voltage when there is a PV power generation. For example, it can be observed that the converters continuously increase its bus voltage during RAS operation because abrupt PV power generation changes its bus voltage and then reduces the rewards in the learning process. By observing the black and green lines in Fig. 9(c) and Fig. 9(d), the low voltage occurred when the centralized reactive power control was implemented. The object functions collided in, thus resulting in a sub-optimal action. In the proposed algorithm, each layer was trained and passes the averaged message over the agent modules between layers. This thus allows the optimal action values with respective object functions to be derived in our framework.

The proposed algorithm separated the controllers into active power and reactive power control to maximize their reward. Thus, rewards r_{y1}^t and r_{y2}^t were implemented to control the reactive power in the converter, whereas reward r_z^t was used to control the active power. However,

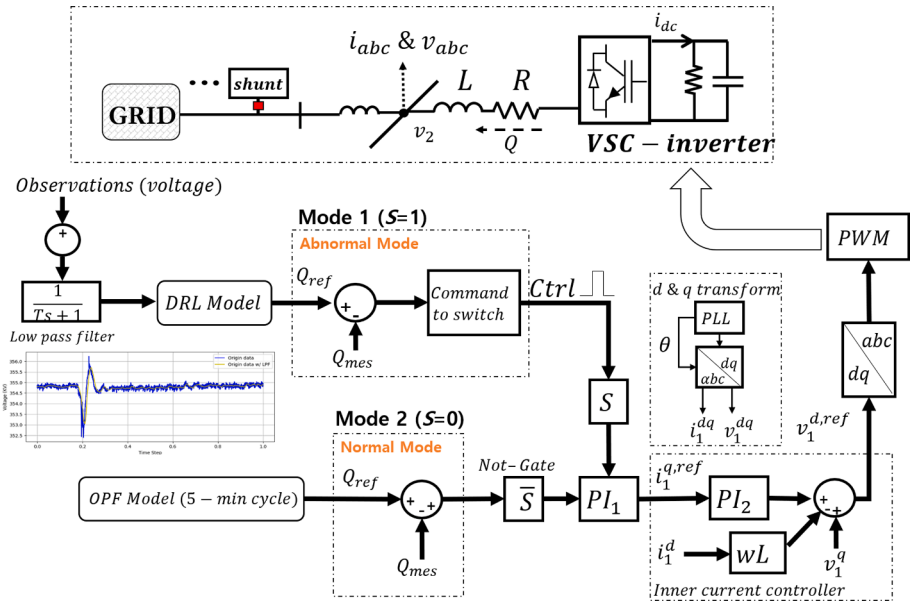


Fig. 5. VSC-HVDC control schemes with low-pass-filter.

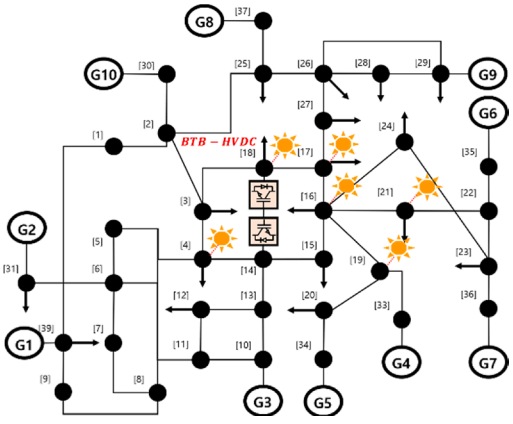


Fig. 6. Modified IEEE-39 standard bus system.

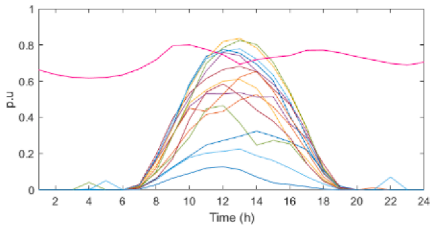


Fig. 7. Example normalized PV and load pattern data.

Table 2
Generator linear cost functions in the 39-bus system.

Gen	p_1	p_2	Gen	p_1	p_2
G1	8.71	0.0031	G6	7.85	0.0019
G2	3.53	0.0074	G7	2.25	0.0014
G3	7.58	0.0066	G8	6.29	0.0041
G4	2.24	0.0063	G9	4.30	0.0051
G5	8.53	0.0069	G10	8.26	0.0032

Table 3
Parameters of PPO model.

Parameters	Values
Batch size	4
Discount factor (gamma)	0.9
Lambda for GAE	0.9
Clip	0.1
Epoch	4
Layer	3
Learning rate for actor, critic Network	0.00002
Optimizer	Adam

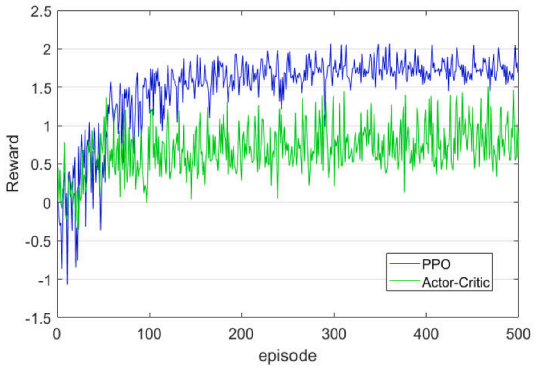


Fig. 8. Reward convergence.

embedded VSC-HVDC only involves line losses using active power control, there is no dramatic difference in total generation cost.

2) CASE 2 - The proposed method was much more applicable to IEEE 33-bus test system. The contingency timeline is the same as the first case. Notable results of this strategy involved the reactive power control from converter. As shown in Fig. 10, the respective bus voltage profiles is obtained for the IEEE-33 bus test system. Frequent voltage deviation (from 1.0 pu) occurred when the actor-critic algorithms was employed, which means that the sub-optimal action can be observed in some simulation cases. However, learning process application to multiple RESs scenarios for given period were all well converged in PPO model. And, fewer deviations were present even under sudden increases in PV

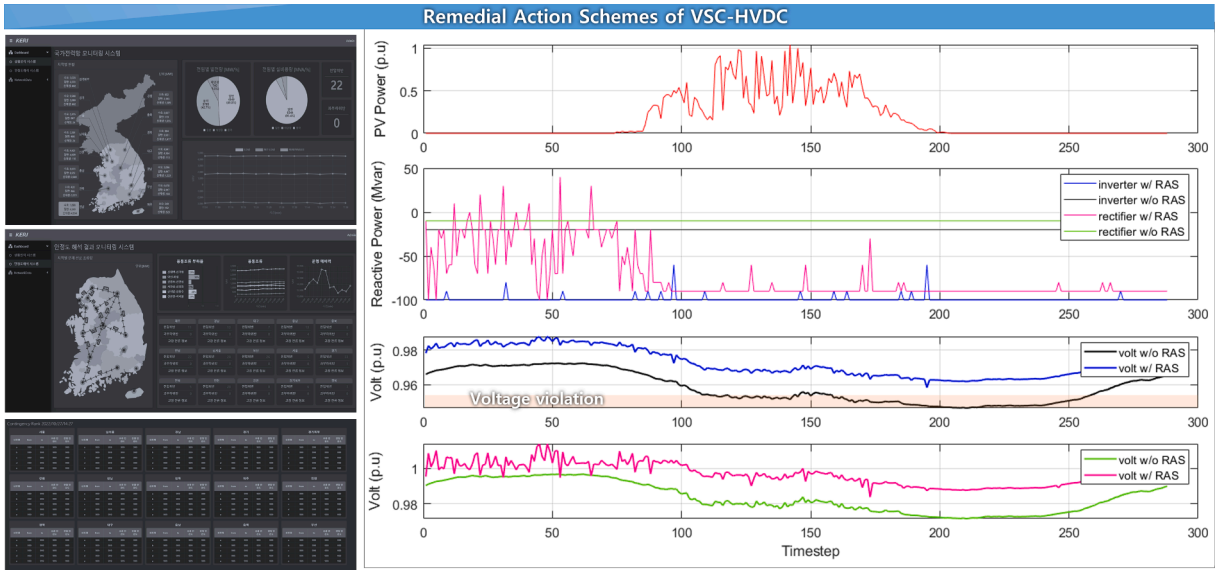


Fig. 9. Action results with proposed RAS platform in Korea Electrotechnology Research institute (a) PV profiles (b) reactive power profiles of converter (c) voltage at 14 bus (d) voltage at 18 bus.

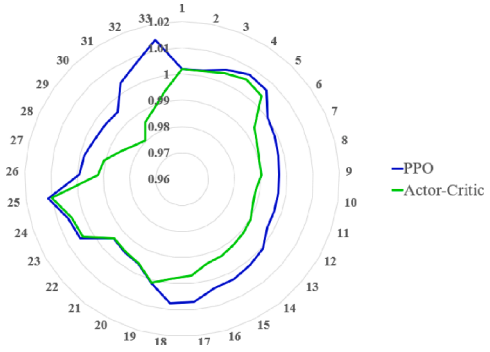


Fig. 10. Bus voltage result (mean value).

power generations, which can lead to instability short- and long-term network operation.

The sudden change of bus voltage from the PV power cannot be regulated when there is a large error between the predicted and real value in EMS. This can be regarded as a normal limitation of the current power network that requires effective RAS to generate an optimal action value in real time. Without the real time control algorithm, the system operator must manage the network stability by activating conventional approaches, such as RESs curtailment. Existing centralized optimal search algorithms such as OPF find possible solutions and then decide upon an optimal one; although they thus can determine the optimal solution, they always require computation time to find suitable value and cannot prevent unexpected network topology changes.

3) CASE 3 - Proposed strategy was also applicable to the Korean power system. The calculation time for stability analysis is performed at 5-minute intervals, and includes overloads, voltage violations, fault currents, and inertia. The contingency timeline is the same as the first case. VSC-HVDC was attached at between *GOYANG* and *JICHUK* bus and

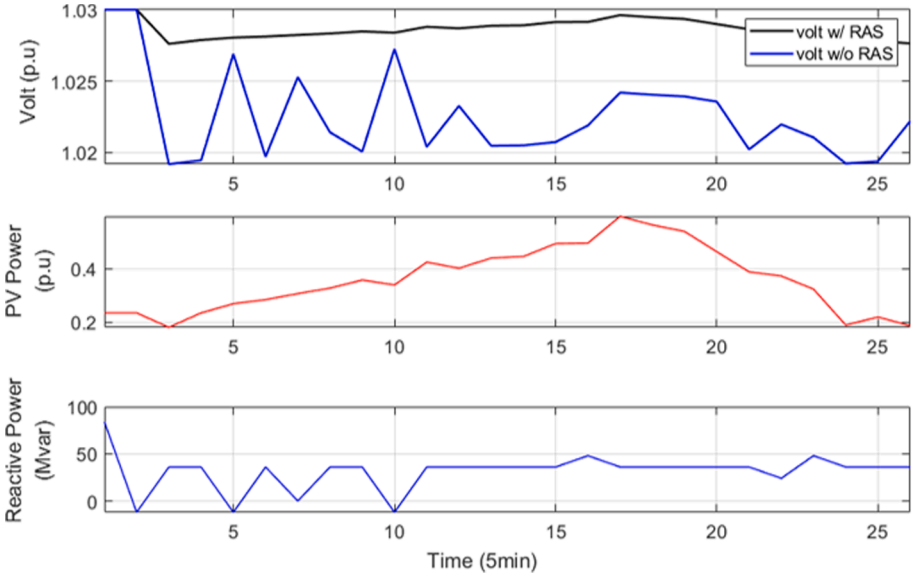


Fig. 11. Action results with proposed RAS platform in Korea power system (a) voltage at *JICHUK* bus (b) PV profiles (c) reactive power profiles of converter.

the converter capacity is 200 MVA. The reward function was formulated for maintaining the voltage as 1.02 pu and it was assumed that three PV plants was operated nearby VSC-HVDC.

Given the increased generation of PV, the converter absorbs a reactive power quickly in abnormal mode, thereby ensuring voltage regulation, as demonstrated by the blue trace in Fig. 11. This ensured that the voltage regulation was sufficient via VSC-HVDC without OPF. However, the voltage profile with the Korean power system is more stable than that with test standard power system because the voltage sensitivity was predominantly affected by total generation capacity; thus, increasing reactive power margin is reasonable to validate results. The total reactive power was easily restricted during the long-term simulation, thereby limiting the capability of the VSC-HVDC as compared to the case where the system size was relatively small. Overall, the proposed algorithm which is synchronized with both EMS and SAS showed that outstanding control performance and autonomously regulated voltage and line flows, thereby validating the effectiveness of the method.

6. Conclusion

A PPO-based RAS framework was proposed to effectively address the voltage issues caused by the high penetration of intermittent RESs. In this paper, BTB VSC-HVDC is targeted and the most stable learning algorithm has been implemented for a large power system. The full observation vector decreases training stability; thus, a PPO algorithm which regulates the size of the trust region in the update has been implemented. To reduce state variable, typical topology configurations are used as inputs by the RL agent to determine the action in the next time step. The learning procedure is stabilized and made robust to continuous changes in network topology. The modeled as intelligent agents have shown a high convergence performance and agent continuously regulates voltage for all time steps in all scenario. And, the frequent change of voltage or reactive power reference can impact the converter stability in weak grid; thus, the proposed control mode is to act as an emergency control with RAS that should be activated in abnormal mode. The results indicate that the real-time HVDC control can be performed in abnormal mode than did the existing EMS method. It can be realized that implementation of an embedded system between real-time digital simulator and server. Its effects on the real power system operation will make for an interesting investigation to be performed addressed in future works. By employing the latest learning model via a cloud service, the proposed optimization model can be kept up-to-date when there is an unexpected power network change.

CRediT authorship contribution statement

Sungyoon Song: Writing – original draft, Conceptualization, Methodology, Software, Visualization. **Yungun Jung:** Investigation, Data curation. **Gilsoo Jang:** . **Seungmin Jung:** Supervision.

Declaration of Competing Interest

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

Data availability

The authors do not have permission to share data.

Appendix

The three linearized state variables of VSC and the state matrix components can be represented by [31]:

$$A_{11} = \left(\frac{k_p^i x_s}{k_p^P x_s i_{q0} - \left(1 + k_p^P\right)}\right)\left(\frac{k_p^P}{1 + k_p^P} - 1\right) i_{q0} - \frac{k_p^i}{1 + k_p^P}$$

$$A_{12} = \frac{\left(\frac{k_p^P}{1+k_p^P} - 1\right) i_{d0} k_v^i x_s}{k_v^p x_s + 1} +$$

$$\left(\frac{-k_p^P k_v^i x_s^2 i_{d0}}{\left(k_v^P x_s + 1\right)\left[k_p^P x_s i_{q0} - \left(1 + k_p^P\right)\right]}\right)\left(\frac{k_p^P}{1 + k_p^P} - 1\right) i_{q0}$$

$$A_{13} = \left(\frac{\left(k_v^P x_s + 1\right)\left(1 + k_p^P\right) v_{sd} + k_p^P x_s i_{d0} v_{sd}}{\left(k_v^p x_s + 1\right)\left[k_p^P x_s i_{q0} - \left(1 + k_p^P\right)\right]}\right)$$

$$\times \left(\frac{k_p^P}{1 + k_p^P} - 1\right) i_{q0} - \frac{\left(\frac{k_p^P}{1+k_p^P} - 1\right) i_{d0} v_{sd}}{k_v^p x_s + 1}$$

$$A_{21} = 0,\ A_{22} = \left(\frac{-k_v^i x_s}{k_v^p x_s + 1}\right), A_{23} = \left(\frac{v_{sd}}{k_v^p x_s + 1}\right)$$

$$A_{31} = k_{PLL}^p \left(\frac{k_P^i x_s}{k_p^P x_s i_{q0} - \left(1 + k_p^P\right)}\right)$$

$$A_{32} = k_{PLL}^p \left(\frac{-k_p^P k_v^i x_s^2 i_{d0}}{\left(k_v^P x_s + 1\right)\left[k_p^P x_s i_{q0} - \left(1 + k_P^p\right)\right]}\right)$$

$$A_{33} = k_{PLL}^{p}\left(\frac{\left(k_v^P x_s + 1\right)\left(1 + k_p^P\right)v_{sd} + k_p^P x_s i_{d0} v_{sd}}{\left(k_v^P x_s + 1\right)\left[k_p^P x_s i_{q0} - \left(1 + k_p^P\right)\right]}\right)$$

References

[1] Alam MS, et al. High-level penetration of renewable energy sources into grid utility: Challenges and solutions. IEEE Access 2020;8:190277–99.
[2] Min D, Ryu J-H, Choi DG. Effects of the move towards renewables on the power system reliability and flexibility in South Korea. Energy Rep 2020;6:406–17.
[3] Yang, H.-T., et al. Placement of fault current limiters in power systems by HFLS sorting and HIGA optimization approach. in Proceedings of the International Conference on Power Systems Transients (IPST2015), Cavtat, Croatia. 2015.
[4] Imhof M, Andersson G. Power system stability control using Voltage Source Converter based HVDC in power systems with a high penetration of Renewables. in 2014 Power Systems Computation Conference. IEEE; 2014.
[5] Cotts BR, Prigmore II JR, Graf KL. HVDC Transmission for Renewable Energy Integration. In: The Power Grid. Elsevier; 2017. p. 171–96.
[6] Gbadamosi SL, Nwulu NI. Reliability assessment of composite generation and transmission expansion planning incorporating renewable energy sources. J Renewable Sustainable Energy 2020;12(2):026301.
[7] Mitra P, Zhang L, Harnefors L. Offshore wind integration to a weak grid by VSC-HVDC links using power-synchronization control: A case study. IEEE Trans Power Delivery 2013;29(1):453–61.
[8] Pan, J., et al. AC grid with embedded VSC-HVDC for secure and efficient power delivery. in 2008 IEEE Energy 2030 Conference. 2008. IEEE.
[9] Li, S., et al. A study on VSC-HVDC based black start compared with traditional black start. in 2009 International conference on sustainable power generation and supply. 2009. IEEE.
[10] Guo Y, et al. Enhanced voltage control of VSC-HVDC-connected offshore wind farms based on model predictive control. IEEE Trans Sustainable Energy 2017;9(1):474–87.
[11] Magdy G, et al. Renewable power systems dynamic security using a new coordination of frequency control strategy based on virtual synchronous generator and digital frequency protection. Int J Electr Power Energy Syst 2019;109:351–68.
[12] Daelemans, G., et al. Minimization of steady-state losses in meshed networks using VSC HVDC. in 2009 IEEE Power & Energy Society General Meeting. 2009. IEEE.
[13] Raza A, et al. Multi-objective optimization of VSC stations in multi-terminal VSC-HVdc grids, based on PSO. IEEE Access 2018;6:62995–3004.
[14] Fu Q, et al. Small-signal stability analysis of a VSC-MTDC system for investigating DC voltage oscillation. IEEE Trans Power Syst 2021;36(6):5081–91.
[15] Luo, F., et al. A Control Strategy of VSC-MTDC Transmission System Based on Voltage Margin. in 2021 IEEE 4th International Electrical and Energy Conference (CIEEC). 2021. IEEE.
[16] Li B, et al. A novel method to determine droop coefficients of DC voltage control for VSC-MTDC system. IEEE Trans Power Delivery 2020;35(5):2196–211.
[17] Wang Y, et al. Adaptive voltage droop method of multiterminal VSC-HVDC systems for DC voltage deviation and power sharing. IEEE Trans Power Delivery 2018;34(1):169–76.
[18] Song S, McCann RA, Jang G. Cost-based adaptive droop control strategy for VSC-MTDC system. IEEE Trans Power Syst 2020;36(1):659–69.
[19] Chavez JCS, et al. A hybrid optimization framework for the non-convex economic dispatch problem via meta-heuristic algorithms. Electr Pow Syst Res 2019;177:105999.
[20] Gonzez AM, Roque AS, Garc-Gonzez J. Modeling and forecasting electricity prices with input/output hidden Markov models. IEEE Trans Power Syst 2005;20(1):13–24.
[21] Ju P, et al. Stochastic dynamic analysis for power systems under uncertain variability. IEEE Trans Power Syst 2017;33(4):3789–99.
[22] Cao D, et al. Attention enabled multi-agent DRL for decentralized volt-VAR control of active distribution system using PV inverters and SVCs. IEEE Trans Sustainable Energy 2021;12(3):1582–92.
[23] Diao, R., et al. Autonomous voltage control for grid operation using deep reinforcement learning. in 2019 IEEE Power & Energy Society General Meeting (PESGM). 2019. IEEE.
[24] Glavic M, Fonteneau R, Ernst D. Reinforcement learning for electric power system decision and control: Past considerations and perspectives. IFAC-PapersOnLine 2017;50(1):6918–27.
[25] Duan J, et al. Deep-reinforcement-learning-based autonomous voltage control for power grid operations. IEEE Trans Power Syst 2019;35(1):814–7.
[26] Wei Q, Liu D, Shi G. A novel dual iterative Q-learning method for optimal battery management in smart residential environments. IEEE Trans Ind Electron 2014;62(4):2509–18.
[27] Li J, et al. Multi-agent deep reinforcement learning for sectional AGC dispatch. IEEE Access 2020;8:158067–81.
[28] Strehl, A.L., et al. PAC model-free reinforcement learning. in Proceedings of the 23rd international conference on Machine learning. 2006.
[29] Bahdanau, D., et al., An actor-critic algorithm for sequence prediction. arXiv preprint arXiv:1607.07086, 2016.
[30] Davari M, Mohamed Y-A-R-I. Robust vector control of a very weak-grid-connected voltage-source converter considering the phase-locked loop dynamics. IEEE Trans Power Electron 2016;32(2):977–94.
[31] Ding, H., et al., Parametric analysis of the stability of VSC-HVDC converters. 2015.
[32] Song H, et al. Optimal electricity supply bidding by Markov decision process. IEEE Trans Power Syst 2000;15(2):618–24.
[33] Schulman, J., et al., Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.
[34] Foerster J, et al. Learning to communicate with deep multi-agent reinforcement learning. Adv Neural Inf Proces Syst 2016;29.
[35] Zimmerman, R.D., C.E. Murillo-Schez, and R.J. Thomas. MATPOWER's extensible optimal power flow architecture. in 2009 IEEE Power & Energy Society General Meeting. 2009. IEEE.