energies

MDPI

Article

Active Exploration by Chance-Constrained Optimization for Voltage Regulation with Reinforcement Learning

Zhenhuan Ding [1], Xiaoge Huang [1] and Zhao Liu [2,*]

[1] Department of Electrical and Computer Engineering, State University of New York at Binghamton, New York, NY 13902, USA; zding1@binghamton.edu (Z.D.); xhuang98@binghamton.edu (X.H.)
[2] School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China
* Correspondence: liuzhao1@bjtu.edu.cn

Abstract: Voltage regulation in distribution networks encounters a challenge of handling uncertainties caused by the high penetration of photovoltaics (PV). This research proposes an active exploration (AE) method based on reinforcement learning (RL) to respond to the uncertainties by regulating the voltage of a distribution network with battery energy storage systems (BESS). The proposed method integrates engineering knowledge to accelerate the training process of RL. The engineering knowledge is the chance-constrained optimization. We formulate the problem in a chance-constrained optimization with a linear load flow approximation. The optimization results are used to guide the action selection of the exploration for improving training efficiency and reducing the conserveness characteristic. The comparison of methods focuses on how BESSs are used, training efficiency, and robustness under varying uncertainties and BESS sizes. We implement the proposed algorithm, a chance-constrained optimization, and a traditional Q-learning in the IEEE 13 Node Test Feeder. Our evaluation shows that the proposed AE method has a better response to the training efficiency compared to traditional Q-learning. Meanwhile, the proposed method has advantages in BESS usage in conserveness compared to the chance-constrained optimization.

Keywords: voltage regulation; chance-constrained optimization; reinforcement learning; uncertainties; battery; active exploration

Citation: Ding, Z.; Huang, X.; Liu, Z. Active Exploration by Chance-Constrained Optimization for Voltage Regulation with Reinforcement Learning. *Energies* **2022**, *15*, 614. https://doi.org/10.3390/en15020614

Academic Editors: Zhengyu Lin, Hui Guo, Fulong Li and Tek Tjing Lie

Received: 6 December 2021
Accepted: 7 January 2022
Published: 16 January 2022

1. Introduction

With increasing PV penetration, power system operation suffers from the uncertainty between a forecast and actual generation. A major research focus is the fact that the voltage of the distribution network becomes vulnerable under the PV impact [1]. BESS is a potential solution to mitigate the effect of the PV uncertainty.

Engineering has a long history of handling uncertainties in application [2–4]. From the mathematical point of view, solutions mitigating the effect of uncertainties can be classified into two major groups. The first group is probability-based optimization. Stochastic programming [5] is a standard algorithm in this group. These techniques are well developed and have been used in real power system applications [6,7]. Chance-constrained optimization, which is a subset of stochastic programming, is a natural framework because it can quantify the probability of the voltage violation [8,9]. While probability-based optimization methods are useful, it is difficult to maintain the balance between its reliability and conservativeness [10], thereby often leading to over-engineered solutions. Conservativeness is necessary for some applications, but sometimes it can increase the control cost of the system [5]. Systems that use probability-based optimization to operate BESS-supplemented grids often need larger than necessary batteries. Another limitation of the probability-based optimization is the model of the uncertainty. The probability-based optimization requires the knowledge of the uncertainty to estimate its distribution or the boundary. In practice, the model of the uncertainty could be inaccurate and hard to derive.

A recent popular group of solutions managing uncertainty is RL. RL was designed to solve the Markov decision process (MDP) which could regard uncertainties without the uncertainty model. The RL agent can grasp the uncertainty features of the historical training data. In voltage control, researchers have already proposed RL methods for different applications. The tap-changer transformer [11], the capacitor-based Volt/Var control [12,13], the load restoration [14], and the reinforcement learning for the autonomous voltage control [15] have been proved to be effective in power systems. RL for BESS is also studied for voltage control [16], but the algorithm proposed in that paper is using the RL to select which BESS should be used instead of directly controlling the charge/discharge power of BESSs. Even though the achievements are made by RL, the long training time or the low training efficiency is always a challenge in real applications [17].

One of the reasons for the low training efficiency of RL is the large exploration process. The exploration strategies of the RL can be categorized as directed and undirected. Exploration is undirected when the action selection is fully random. Typical exploration methods, such as ε-greedy, Boltzmann distribution, and SoftMax [18], are undirected exploration strategies. The lack of the utilization of specific information characterizes the undirected exploration strategy, in which RL agent are randomly selects action choices with uniform probability for exploration. The directed exploration methods take advantage of specific information or knowledge to guide the exploration. The active exploration (AE) method was first proposed in [19], whereby human knowledge is used to correct the improper actions during the training process of an inverted pendulum system. Such a knowledge-based exploration method is also developed for guiding the RL agent in a maze environment by modification of the activation function [20]. A dynamic parameter tuning for exploration of a meta-learning method with a human interaction robot system is verified in [21]. The parameter changes the probability distribution of the action selection, which avoids the ineffective training. In a recent study, the AE concept is applied to controlling an HVAC system by a deep Q-learning method to accelerate the training speed [22]. The authors use engineering knowledge to judge the action picked by RL agents to achieve faster convergence purpose. Traditional reinforcement learning agents adopt a simple strategy, by randomly selecting an action from a set with all the available actions, which wastes training time significantly. The above studies point out that if the human knowledge or engineering sense can be utilized to guide the action selection during the training process, then the overall training time reduces.

The existing literature shows the potential of probability-based optimization and RL methods in obtaining better performance of voltage regulation with renewable uncertainties. However, answers to the following three issues are still unsettled, which motivates this paper's research:

(1) The forecasting model of the renewable generation is hard to obtain accurately. The mismatch between the predicted and actual value diminishes the performance of the probability-based optimization. Therefore, how to overcome the effect of the forecasting mismatch is the challenge of the probability-based optimization methods.
(2) The conserveness characteristic of the probability-based optimization methods requires a larger size of the BESS to compensate for the forecasting mismatch. Knowing how to reduce the conserveness of the traditional methods so that the BESS size can be managed properly is valuable for industries.
(3) RL has advantages in handling uncertainties. However, the low training efficiency of RL limits its application. Utilizing the probability-based optimization to improve the training time is meaningful.

In this paper, we integrate the conventional engineering knowledge with a relatively novel RL to conquer the aforementioned issues. Specifically, the chance-constrained optimization is employed as the engineering knowledge of proposed AE methods to improve the training efficiency in voltage regulation problems, and the RL framework relieves the dependency on the forecasting accuracy and the conserveness. The contributions of the work are concluded as:

(1) We propose an AE method for the voltage regulation problem by applying BESSs in an RL framework. The performance of the proposed method is verified in an IEEE standard test feeder. Simulation results reveal that the proposed method has a better performance.
(2) The proposed method speeds up the training process by modifying the action selection distribution according to the engineering knowledge, which discourages unnecessary exploration. To validate the improvement in the training efficiency, the proposed method is compared with the conventional Q-learning method as a benchmark method.
(3) The effectiveness of the BESSs' usage is improved by the proposed method. By comparing the proposed method with a chance-constrained optimization, the proposed method can achieve the voltage regulation with a smaller BESS size while the chance-constrained optimization method returns infeasible solution.

The rest of the paper is organized as follows. Section 2 demonstrates the chance-constrained optimization, which is regarded as the engineering knowledge. A conventional Q-learning and our proposed method is given in Section 3. Section 4 illustrates the case studies. A conclusion is drawn in Section 5.

2. Engineering Knowledge of Voltage Regulation

One of the limits of the RL is the low exploration efficiency. In this section, an active exploration approach is introduced to enhance exploration efficiency. To elaborate, we engage engineering knowledge to guide the exploration process by the action selection probability distribution. A chance-constrained optimization is firstly applied to obtain an optimal operation profile for BESSs. Then, the obtained optimal BESSs' profile will guide the action selection in the exploration phases.

2.1. Forecasting Mismatch Model for Uncertainty Distribution

The initial process of the chance-constrained optimization is to obtain a mismatch uncertainty distribution, which is determined by different weather conditions, as Figure 1 shows. The forecasting mismatch refers to the difference between the forecasted power and the actual PV generation. The PV forecasting is sensitive to the weather condition. According to our previous study [23], the accuracy of the forecasting depends on the weather condition. For example, the mismatch between the PV prediction and the actual generation in cloudy days is larger than the sunny and overcast days. Therefore, the model of the mismatch uncertainty depends on the weather conditions. Figure 1 gives the probability density function (PDF) of the forecasting mismatch between the actual PV generation and the forecasted generation under various weather conditions. The PDF is the regression results from the original forecasting mismatch data, which describes the average mismatch of a day. We record the average mismatch of 138 days with the classification of weather conditions and obtained different PDFs of the forecast mismatch. The forecasting algorithm has better performance in accuracy under the weather of the clear and overcast skies than cloudy condition. Therefore, voltage regulation algorithms would encounter a different uncertainty set in this study.

2.2. Voltage Regulation with Chance-Constrained Optimization

The goal we set for the voltage regulation algorithm is to regulate the feeder voltage and minimize the BESS operation cost. The cost is modelled as the throughput degradation cost of batteries [24].

$$\min C(\boldsymbol{P_{BESS}}), \tag{1}$$

Subject to

$$SOC_{min} \leq SOC_{k,t-1} + \Delta t \cdot \eta_k P_{BESS,k,t} \leq SOC_{max}, \tag{2}$$

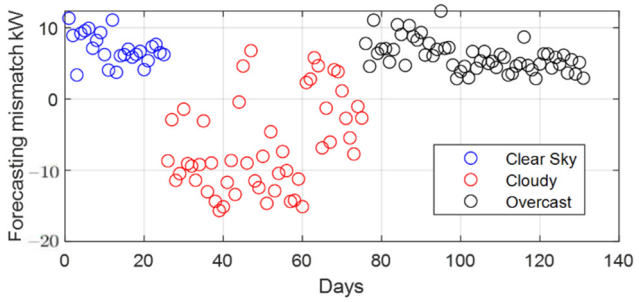
$$\boldsymbol{V_{min}} \leq Loadflow(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{P_{BESS}}) \leq \boldsymbol{V_{max}}, \tag{3}$$

$$\boldsymbol{P_{min}} \leq \boldsymbol{P_{BESS}} \leq \boldsymbol{P_{max}}, \tag{4}$$

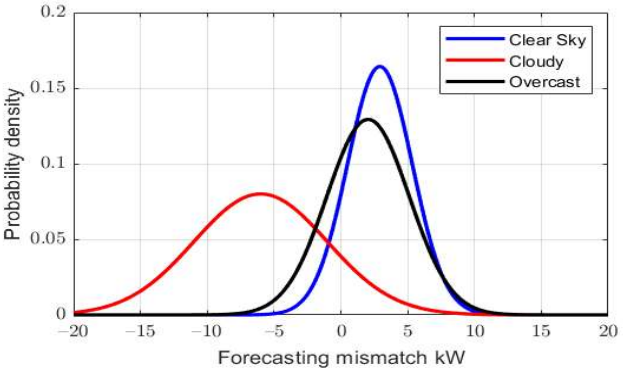
where $\boldsymbol{P_{BESS}}$ denotes a vector of the BESS operation profile, which is the optimization variable. The element in the vector is $P_{BESS,k,t}$. $P_{BESS,k,t}$ denotes the kth BESS operating power at time slot t. $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the forecasted real and reactive power. η_k is the charge or discharge efficiency of the BESS. $\eta_k = \eta_c$ if the BESS is charging and $\eta_k = 1/\eta_d$ if the BESS is discharging. η_c denotes the charging efficiency and η_d refers to the discharging efficiency. The SOC of BESS updates according to the $P_{BESS,k,t}$. Δt is a 15-min operation interval. SOC_{min} and SOC_{max} are the minimum and maximum SOC values. $Loadflow(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{P_{BESS}})$ means the load flow calculation, which outputs voltages. The $\boldsymbol{V_{min}}$ and $\boldsymbol{V_{max}}$ are vectors containing the minimum voltage and maximum voltage values. The operation range of the BESS is from -200 kW to 200 kW, which determines the vectors of $\boldsymbol{P_{min}}$ and $\boldsymbol{P_{max}}$. $C(\boldsymbol{P_{BESS}})$ is the degradation of the BESS. It is linear with the $\boldsymbol{P_{BESS}}$.

$$C(\boldsymbol{P_{BESS}}) = \sum_{t=1}^{T} \frac{C_{bat}}{2 \cdot L_{bat}(DOD) \cdot DOD} \cdot \left| P_{BESS,k,t} \right| \cdot \Delta t,$$

where C_{bat} denotes the battery capital cost, $L_{bat}(DOD)$ determines the battery life coefficient by considering the depth of discharge (DOD). C_{bat}, DOD and $L_{bat}(DOD)$ are predefined values.



(**a**)



(**b**)

Figure 1. Uncertainty model of the forecasting mismatch: (**a**) daily average forecasting mismatch data; (**b**) regression of the forecasting mismatch by weather conditions.

The voltage in the distribution network changes with the load and PV. A load flow calculation is necessary to obtain the accurate voltage value of a specific node. Many researchers agree that the voltage can be linearly expressed by the injected power [9]

$$\boldsymbol{V} = \boldsymbol{RP} + \boldsymbol{XQ} + \boldsymbol{V}_0, \tag{5}$$

where $\boldsymbol{V} = [V_1, \ldots, V_N]^{\boldsymbol{T}}$ denotes the voltage profile of the whole feeder. $\boldsymbol{R}, \boldsymbol{X} \in \mathbb{R}^{N \times N}$ are matrices containing the elements R_{ij} and X_{ij} which are in the ith row and the jth column, respectively. Note that $\boldsymbol{R}$ and $\boldsymbol{X}$ is different from the reactance and inductance of the system

impedance. $\boldsymbol{R} = \frac{\partial V}{\partial P}$, and $\boldsymbol{X} = \frac{\partial V}{\partial Q}$. N is the total number of phases. $\boldsymbol{P} = [P_1, \ldots, P_N]^T$ and $\boldsymbol{Q} = [Q_1, \ldots, Q_N]^T$ are the active power and reactive power injections. $\boldsymbol{V}_0$ is the reference voltage that can be obtained from the default net power.

Because the active power has a major impact on the voltage magnitude under high $\frac{R}{X}$ ratio distribution networks, only the active power injection is considered. Therefore, the voltage variation is:

$$\Delta \boldsymbol{V} = \boldsymbol{R}\Delta \boldsymbol{P}, \tag{6}$$

where $\Delta \boldsymbol{V}$ is a vector containing the voltage variation caused by the fluctuation of the active power injections $\Delta \boldsymbol{P}$.

When the forecasting mismatch and the linearization are introduced, the (1c) are modified as:

$$\boldsymbol{V_{min}} \leq \boldsymbol{R}(\boldsymbol{P} + \boldsymbol{P_{BESS}} + \boldsymbol{\xi}) + \boldsymbol{XQ} + \boldsymbol{V}_0 \leq \boldsymbol{V_{max}}, \tag{7}$$

where $\boldsymbol{\xi}$ is the uncertainty of the net power, which can be derived from Section 2.1.

Equation (1) can be formulated in a probabilistic fashion.

$$\min C(\boldsymbol{P_{BESS}}), \tag{8}$$

Subject to

$$SOC_{min} \leq SOC_{k,t-1} + \Delta t \cdot \eta_k P_{bat,k,t} \leq SOC_{max}, \tag{9}$$

$$\mathbb{P}(\boldsymbol{V_{min}} \leq \boldsymbol{R}(\boldsymbol{P} + \boldsymbol{P_{BESS}} + \boldsymbol{\xi}) + \boldsymbol{XQ} + V_0) \geq \alpha, \tag{10}$$

$$\mathbb{P}(\boldsymbol{R}(\boldsymbol{P} + \boldsymbol{P_{BESS}} + \boldsymbol{\xi}) + \boldsymbol{XQ} + V_0 \leq \boldsymbol{V_{max}}) \geq \alpha, \tag{11}$$

$$\boldsymbol{P_{min}} \leq \boldsymbol{P_{BESS}} \leq \boldsymbol{P_{max}}, \tag{12}$$

where $\mathbb{P}(\cdot)$ is the probability of the event. Equation (10) and Equation (11) mean the probability that voltage is within the designed range in every time slot t is more than or equal to α. We use the individual constrained format. In this work, the α is set to 0.95.

Define,

$$g_1(\boldsymbol{P_{BESS}}, \boldsymbol{\xi}) = \boldsymbol{R}(\boldsymbol{P} + \boldsymbol{P_{BESS}} + \boldsymbol{\xi}) + \boldsymbol{XQ} + V_0 - V_{min}, \tag{13}$$

$$g_2(\boldsymbol{P_{BESS}}, \boldsymbol{\xi}) = \boldsymbol{V_{max}} - \boldsymbol{R}(\boldsymbol{P} + \boldsymbol{P_{BESS}} + \boldsymbol{\xi}) + \boldsymbol{XQ} + V_0. \tag{14}$$

The Formulations (10) and (11) can be rewritten as:

$$\mathbb{P}(g_1(\boldsymbol{P_{BESS}}, \boldsymbol{\xi}) \geq 0) \geq \alpha, \tag{15}$$

$$\mathbb{P}(g_2(\boldsymbol{P_{BESS}}, \boldsymbol{\xi}) \geq 0) \geq \alpha. \tag{16}$$

In our formulation, the voltage approximation is linear, so that Equations (15) and (16) have a decoupled format to $\boldsymbol{P_{BESS}}$ and $\boldsymbol{\xi}$.

$$g_1(\boldsymbol{P_{BESS}}, \boldsymbol{\xi}) = h_1(\boldsymbol{P_{BESS}}) - k_1(\boldsymbol{\xi}), \tag{17}$$

$$g_2(\boldsymbol{P_{BESS}}, \boldsymbol{\xi}) = h_2(\boldsymbol{P_{BESS}}) - k_2(\boldsymbol{\xi}), \tag{18}$$

where,

$$\begin{gathered} h_1(\boldsymbol{P_{BESS}}) = \boldsymbol{R}(\boldsymbol{P} + \boldsymbol{P_{BESS}}) + \boldsymbol{XQ} + V_0 - \boldsymbol{V_{min}} \\ k_1(\boldsymbol{\xi}) = -\boldsymbol{R\xi} \\ h_2(\boldsymbol{P_{BESS}}) = \boldsymbol{V_{max}} - \boldsymbol{R}(\boldsymbol{P} + \boldsymbol{P_{BESS}}) + \boldsymbol{XQ} + V_0 \\ k_2(\boldsymbol{\xi}) = \boldsymbol{R\xi} \end{gathered}$$

Then the chance constraints (15) and (16) get numerically simplified to linear [23].

$$\mathbb{P}(g_1(\boldsymbol{P_{BESS}}, \boldsymbol{\xi}) \geq 0) \geq \alpha \;\Leftrightarrow h_1(\boldsymbol{P_{BESS}}) \geq q_\alpha(k_1(\boldsymbol{\xi})), \tag{19}$$

$$\mathbb{P}(g_2(\boldsymbol{P_{BESS}}, \boldsymbol{\xi}) \geq 0) \geq \alpha \;\Leftrightarrow h_2(\boldsymbol{P_{BESS}}) \geq q_\alpha(k_2(\boldsymbol{\xi})), \tag{20}$$

where $q_\alpha(k(\boldsymbol{\xi}))$ is the α-quantile of $k(\boldsymbol{\xi})$. At this point, the uncertainties are formulated linearly.

Then the overall chance-constrained optimization formulation will be convex.

$$\min C(\boldsymbol{P_{BESS}}), \tag{21}$$

Subject to:

$$SOC_{min} \leq SOC_{k,t-1} + \Delta t \cdot \eta_k P_{bat,k,t} \leq SOC_{max}, \tag{22}$$

$$h_1(\boldsymbol{P_{BESS}}) \geq q_\alpha(k(\boldsymbol{\xi})), \tag{23}$$

$$h_2(\boldsymbol{P_{BESS}}) \geq q_\alpha(k(\boldsymbol{\xi})), \tag{24}$$

$$\boldsymbol{P_{min}} \leq \boldsymbol{P_{BESS}} \leq \boldsymbol{P_{max}}. \tag{25}$$

The optimization solver used in the work is the *cvx_toolbox* published by Stanford University [25].

3. Proposed Method: Battery Energy Storage Systems (BESS) Operation by an Active Exploration (AE) Reinforcement Learning

The Markov decision process is the base of the RL. In this section, we discuss the MDP formulation of the BESS operation. Then, a modified AE Q-learning framework is proposed to solve the problem.

3.1. Markov Decision Process (MDP) of BESS Operation

MDP is with a 4-tuple of S, A, Pr_a, and R_a. S denotes the state space where $s \in S$. A represents the action space where $a \in A$ The transition probability is $Pr_a = Pr_a(s_{t+1}|a, s_t)$. R_a is the immediate reward obtained after transitioning from state s_t to state s_{t+1} with the action a.

State: In our problem, the first state is the voltage magnitude of the node per unit.

$$s_v = \begin{cases} 1, \ if \ v < 0.945 \\ 2, \ if \ 0.945 \leq v \leq 0.95 \\ 3, \ if \ 0.95 \leq v \leq 1.05 \\ 4, \ if \ 1.05 \leq v \leq 1.055 \\ 5, \ if \ 1.055 < v \end{cases}, \tag{26}$$

The second state is the SOC of the BESS. The range of the SOC is from 0 to 1. The interval between each SOC state is determined by the value of the charging or discharging power of the action space. The total state will be the product of the voltage state and the SOC state, $\boldsymbol{S} = \boldsymbol{S_v} \times \boldsymbol{S_{SOC}}$.

Action: Given a state S, a BESS executes the action a_{BESS} at the time slot t. The a_{BESS} represents the charging or discharging power. To avoid violating the power constraints, the action is within the minimum and maximum power as follows:

$$a_{BESS} \in \left[P_{b,min}, P_{b,min} + \Delta^b, P_{b,min} + 2\Delta^b, \ldots, P_{b,max}\right], \tag{27}$$

where $\Delta^b = \frac{P_{b,max} - P_{b,min}}{N^b}$ is used to discretize the action set with N^b which is the number of the actions for a single BESS. In this work, N^b is set to 21 and $P_{b,min}$ is -200 kW and $P_{b,max}$ is 200 kW. Δ^b is 20 kW. This range is the same as the chance-constrained optimization.

Transition probability: the state space contains two terms, the voltage and SOC. For the voltage of the feeder, it is updated with the change of net power injected. For the state of SOC, it is determined by the previous state and actions.

$$S_{SOC,t+1} = S_{SOC,t} + \eta a_{BESS,t} \Delta t, \tag{28}$$

where η is the charge/discharge efficiency. $\eta = \eta_c$, if $a_{BESS,t}$ is positive, which means the BESS is charging. $\eta = 1/\eta_d$, if $a_{BESS,t}$ is negative, which means the BESS is discharging. η_c and η_d refer to the charging and discharging efficiency. Δt is the time interval between each action. Our problem sets 15 min as the time interval. Therefore, the total number of steps for one day is 96.

Reward: The principle of the reward design is to encourage the proper actions and discourage improper actions. The total reward consists of different categories: voltage, SOC and degradation. The following is a detailed explanation of the rewards.

For voltage reward, if the selected action makes the voltage magnitudes hold within the safety range of [0.95, 1.05] in p.u., then the reward for the voltage is designed to be 1. If the action causes any voltage violation, the punishment is set to -100. The punishment should have a greater absolute value than the reward because the total steps in one episode are 96. If the punishment is smaller than 96, the training process may converge to a local optimal.

$$R_{v,t} = \begin{cases} 1, \ if \ v \in [V_{min}, V_{max}] \\ -100, \ else \end{cases}. \tag{29}$$

The reward of the SOC follows the similar logic: if the BESS's SOC is in $[SOC_{min}, SOC_{max}]$, the reward is 1. Otherwise, the punishment is -100.

$$R_{SOC,t} = \begin{cases} 1, \ if \ SOC \in [SOC_{min}, SOC_{max}] \\ -100, \ else \end{cases}. \tag{30}$$

The reward of the degradation cost is the same as the chance-constrained optimization. The Q-learning looks for the maximum value of the sum of the reward. The degradation reward is negative value.

$$R_{D,t} = -k|a_{BESS}|, \tag{31}$$

where k is a constant related to the degradation coefficient of BESS as Equation (1).

The overall immediate reward is a weighted combination of each individual reward.

$$r_t = \theta_1 R_{v,t} + \theta_2 R_{SOC,t} + \theta_3 R_{D,t}, \tag{32}$$

where θ_1, θ_2 and θ_3 are coefficients of the reward.

3.2. Active Exploration (AE) during the Training

The final product of Q-learning is an optimal policy π^* which maximizes the $Q_{\pi^*}(s_t, a_t)$ such that:

$$Q_{\pi^*}(s_t, a_t) = \max_{\pi} Q_{\pi}(s_t, a_t), \tag{33}$$

where $Q_{\pi}(s_t, a_t)$ denotes the grade of selecting the action a_t at state s_t. π is the policy determining the action of BESSs under a certain state. According to the observed the state, the traditional RL agent selects actions based on the ε-greedy algorithms [26].

$$a_t = \begin{cases} argmax_{a_t} Q(s_t, a_t), \ with \ probability \ 1 - \varepsilon \\ random \ a_t, \ with \ probability \ \varepsilon \end{cases}. \tag{34}$$

By contrast with the traditional action selection approach, modifications are introduced to improve the training efficiency. The conventional RL processes the exploration by assuming the action selection follows a distributed uniform distribution. The RL agent has N^b different action that can be taken, and each possible action has equal probability. Then, the exploration is with probability as follows,

$$\mathbb{P}(P_{b,min} \leq a_t \leq P_{b,max}) = \frac{1}{N^b}, \tag{35}$$

where $\mathbb{P}(\bullet)$ is the probability distribution function.

After the optimization problem (10) is solved, the optimal BESS operation result $P^*_{bat}(t)$ is known. The action a_t can be selected based on that value. In this study, we regard the $P^*_{bat}(t)$ as the engineering knowledge. The engineering knowledge or experience can accelerate the training process of active exploration. The chance-constraint optimization in our study can only derive a relatively conservative solution. The RL method is supposed to be less conservative because RL converges to an average optimal instead of worst-case satisfactory.

In this study, we define the active exploration that the action selection is actively assumed to follow a Gaussian distribution during the exploration.

$$\mathbb{P}(P_{b,min} \leq a_t \leq P_{b,max}) = \frac{1}{\sigma\sqrt{2\pi}} \int_{P_{b,min}}^{P_{b,max}} e^{-\frac{(a_t-\mu)^2}{2\sigma^2}} da_t, \tag{36}$$

where σ^2 is the variance of the distribution and μ is the mean of distribution. In our active exploration, $\mu = P^*_{bat}(t)$, so that the RL agent has a higher probability to select a rational action during the training phase, which results in the improvement of the convergence rate and the accuracy. The variance should be designed according to the confidence level of the engineering knowledge. If the forecasting algorithm performs well in terms of accuracy, then the chance-constrained optimization could generate reliable results. The σ could be given as a relatively large value. Otherwise, σ can be a small value to extend the exploration area and corresponding probability. As a trade-off, the smaller σ results in a longer training period.

The comparison between the active exploration and the conventional exploration is illustrated in Figure 2. Compared to the uniform distribution exploration used by the conventional reinforcement, the probability that an RL agent chooses the action that is close to the $P^*_{bat}(t)$ is higher under the Gaussian distribution exploration. The corresponding reward of the active exploration is supposed to be larger than the conventional uniform distribution because of the engineering knowledge. The larger reward can be regarded as positive feedback to accelerate the speed of the learning.

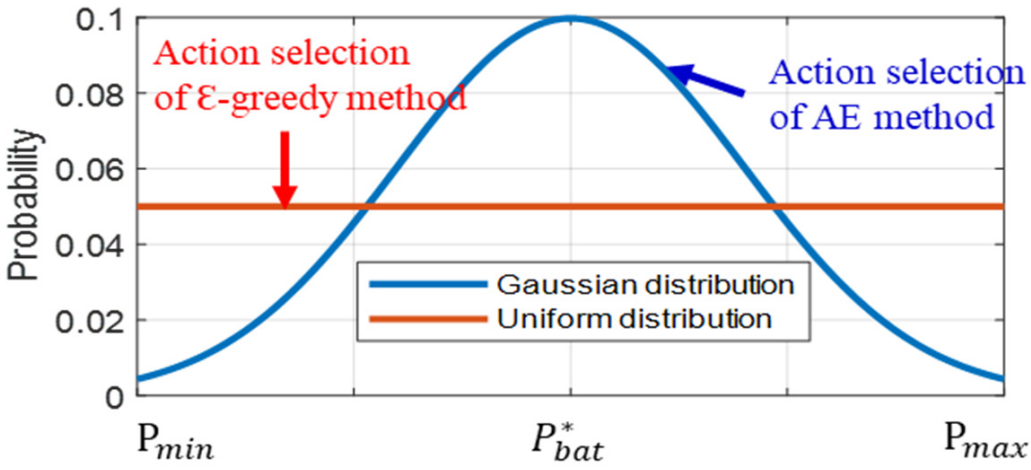


Figure 2. Probability density function of the action selection.

After the actions of BESSs are selected, the environment interacts with the actions and generates the immediate rewards. The environment is the IEEE 13 Node Test Feeder in the OpenDSS [27]. The implementation of the RL is demonstrated in Figure 3.

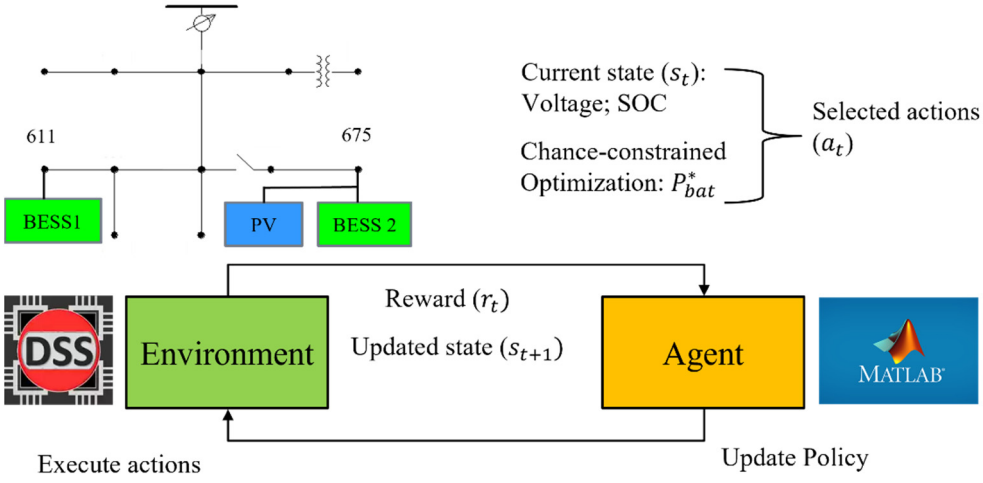


Figure 3. Proposed method implementation.

By contrast with the optimization method, the RL method does not require the linear approximation of the voltage variation. The function of voltage to the injected power is a non-linear and even non-convex mapping, which ensures the accuracy of the power flow.

$$\boldsymbol{V} = loadflow(\boldsymbol{P_t}, \boldsymbol{Q_t}, \boldsymbol{a_t}), \tag{37}$$

The unbalanced load flow calculation algorithm is backward forward swept (BFS) algorithm in the OpenDSS software. The SOC update refers to (3).

When the action a_t is executed, the reward $r_t(s_t, a_t)$ in (17) is calculated and the next state s_{t+1} is evaluated. The rule for the update of the action-value function $Q(s_t, a_t)$ is [26].

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \beta\left[r(s_t, a_t) + \gamma \cdot \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\right], \tag{38}$$

where β and γ stand for the learning rate and discount rate. a' denotes all potential actions. The Q-learning update (23) takes place in the Matlab.

4. Case Study and Results

We evaluate the proposed algorithm in four aspects: (1) voltage regulation performance; (2) training efficiency improvement; (3) conserveness; (4) success rate of the voltage regulation under different uncertainty levels and different BESS sizes. Specifically, the feasibility of proposed method in voltage regulation is given in Section 4.1. The proposed AE method is compared with the conventional RL to demonstrate the training efficiency improvement in Section 4.2. The conserveness study between the proposed method and the chance-constrained optimization occurs in Section 4.3. In the last section, the success rates of the proposed method, the conventional Q-learning, and the chance constrained optimization are studied to reveal their performance under different weather conditions.

Two BESSs are placed in the IEEE 13 Node Test Feeder. BESSs are installed in node 611 and node 675. Node 611 is the location with the most vulnerable node to the undervoltage problem. Node 675 is the most possible location of the overvoltage problem. The rest of the feeder maintains the default values. In our experiment, the tap position of the transformer remains unchanged. The power capacity of each BESS is 200 kW. The individual robust parameter α is set to 0.99. For RL, the learning rate β is 0.8 and the discount factor γ is set to be 0.5. The aforementioned parameters are tuned based on the training performance to the historical data. Different parameters will result in different performance in training speed, but the reinforcement learning agent will converge to the same optimal policy with different learning and discount rates.

4.1. Voltage Regulation Results of the Proposed Method

In this section, the feasibility of the proposed method is tested. We use the 300 days of the load and PV data as input, and examine the proposed method in the case of exterior day. The energy capacity is 400 kWh for each BESS.

Figure 4 shows the results of the voltage regulation, and it illustrates the voltage profile for algorithms. In the voltage regulation results, the blue lines denote the voltage profiles of node 611, and the red lines denote the result in node 675. The x-axis is the hour of the day, and the y-axis denotes the voltage magnitude in per unit. The minimal allowed voltage in the out formulation is 0.95 p.u. and the maximum voltage limit is 1.05 p.u. The thin curves are the original voltage profiles. The bold lines are the results of the proposed method. We can see that, after the RL agent executes, the voltage profiles are managed in the range of [0.95, 1.05] p.u.

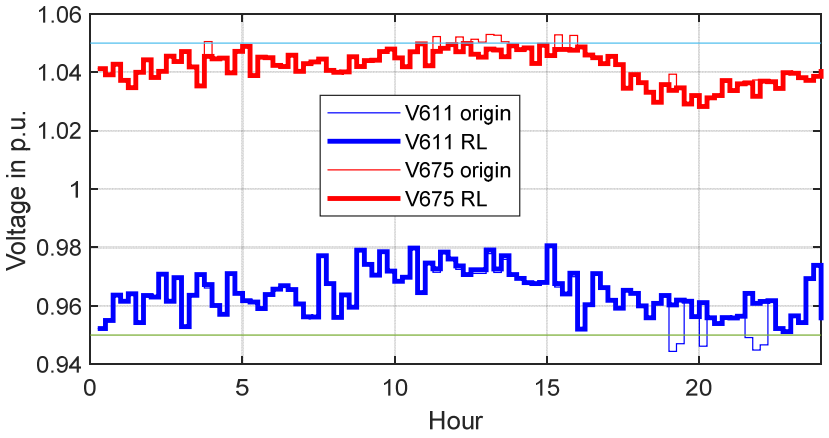


Figure 4. Voltage regulation results.

Undervoltage violations happen in node 611 after sunset. Overvoltage violations take place in node 675 at noon. The bold curves represent the voltage results from the proposed method. The red curve shows the voltage results from our method. Overvoltage and undervoltage issues caused by a PV are resolved by the proposed algorithm.

In Figure 5, the operation power and SOC of two BESSs are given. The x-axis is the hour of the day. The curves with cycles represent the operation power in kW, and the rest of the lines denote the state of charge of the BESSs. From the simulation, we can conclude that the BESSs run within the designed SOC constraints. The feasibility of the proposed method is verified.

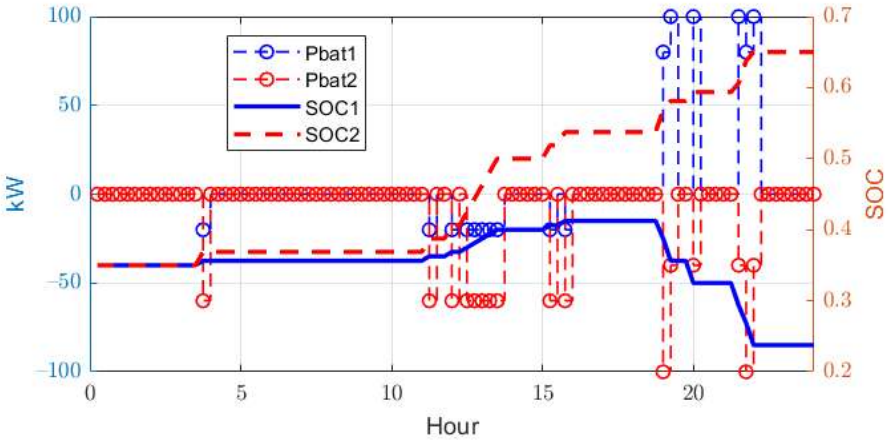
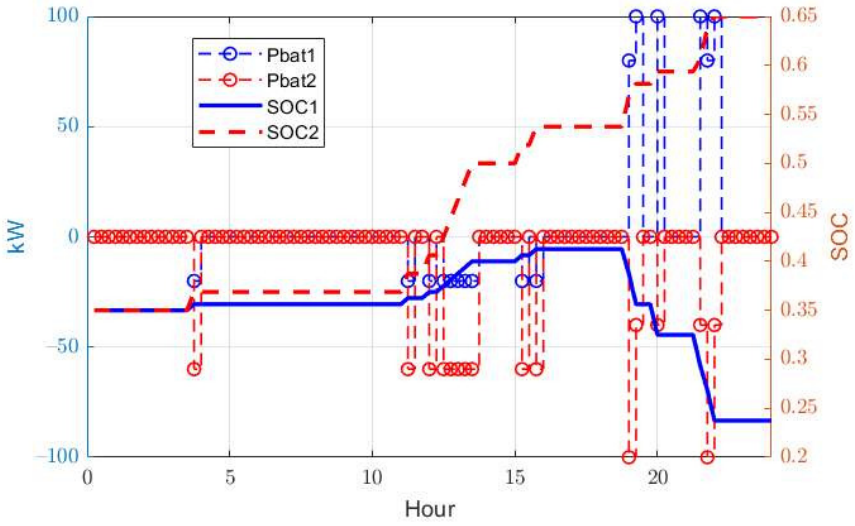


Figure 5. Battery energy storage systems (BESSs) operation results.

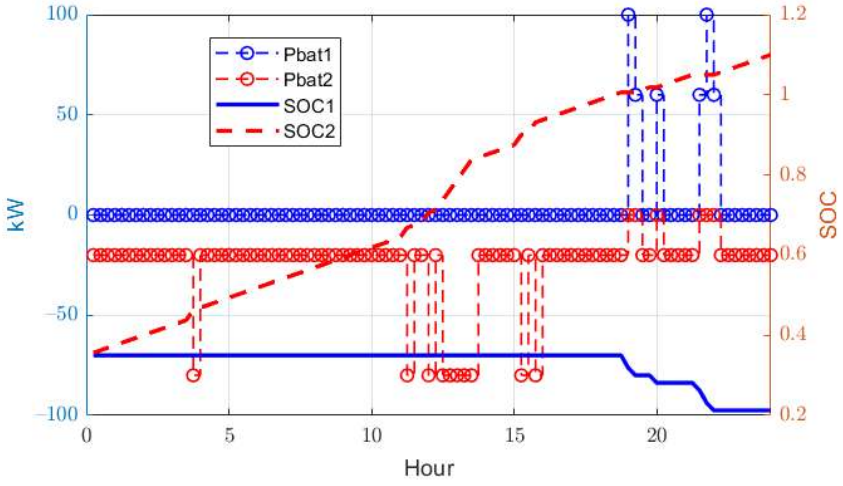
4.2. Compare with Conventional Q-Learning by the Training Efficiency

The feasibility and the training efficiency of the proposed method and the conventional Q-learning (QL) are tested in this section. We use 100 days of the load and PV data as inputs, and examine the proposed method in the case of exterior day for the feasibility test. We would like to show our proposed method has advantages in training efficiency compared to QL.

In Figure 6, the operation power and SOC of two BESSs are given. The maximum of the training episode is set to 3000 in this case. The x-axis is the hour of the day. The left y-axis is the operating power of BESS in kW. The right y-axis denotes the SOC. From the simulation with the 3000 episodes, we can conclude that the BESSs run within the designed constraints of the proposed AE method. However, the SOC of QL is out of the limits of [0, 1]. That means the proposed method converges to an optimal policy at 3000 episodes while the conventional QL is infeasible at the same training efforts, which indicates the improvement of the proposed AE method.



(**a**)



(**b**)

Figure 6. BESSs operation results under the maximum training episode of 3000: (**a**) Power profile and the state of charge (SOC) of the active exploration (AE) method; (**b**) power profile and the SOC of the Q-learning (QL) method.

Figure 7 gives the accumulated rewards and the average values of QL and AE. The maximum of the training episode is still set to 3000. The x-axis is the number of the episode. Y-axis is the reward value. According to the reward formulation, if the final reward value is larger than zero, then no voltage violation and all operation constraints are satisfied. If the final reward value is smaller than zero, that means the constraints are violated. Both the reward and the average reward of the proposed method is higher than the corresponding QL. At the end of the training, the reward value of the proposed method is positive, but the QL is negative. From the numerical simulation, we can conclude that the proposed AE method accelerates the training process comparing to the conventional QL.

Figure 8 gives the success rates of QL and AE. The maximum of the training episode is set from 1000 to 6000 in this case. We use 300 days of the load and PV data as inputs, and exam the proposed method in the case of 100 different days.

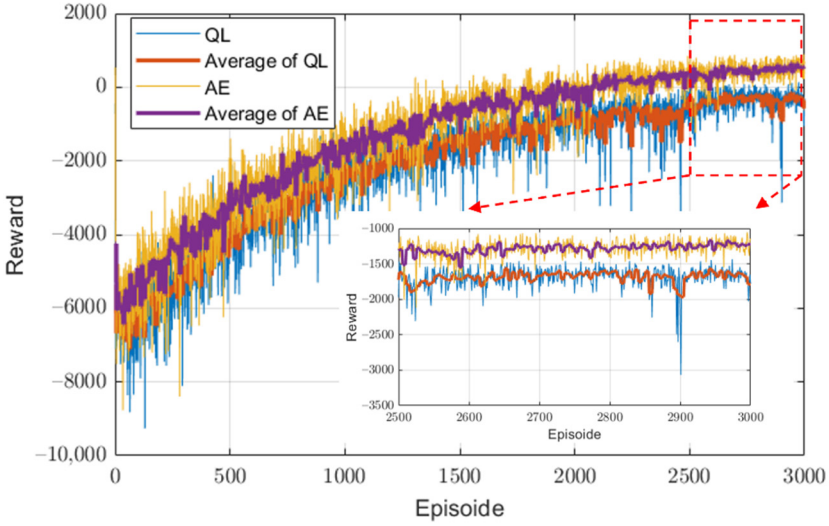


Figure 7. Rewards and average rewards of QL and AE.

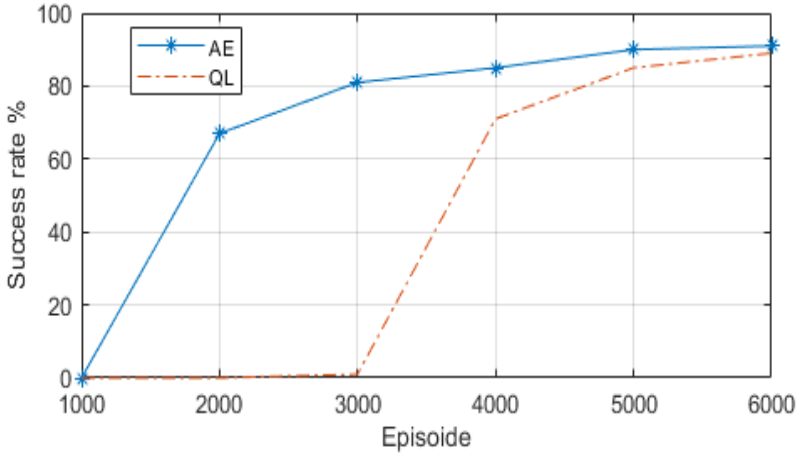


Figure 8. Rewards and average rewards of QL and AE.

The x-axis is the number of the episode. Y-axis is the success rate. In a specific test profile, the OpenDSS calculates the voltage magnitudes. MATLAB updates the SOC of all BESS according to (13). We set the voltage allowed range as [0.95, 1.05] and the SOC range as [0, 1]. If the test results tell us that all the voltages and SOCs are within the allowed range, we define the corresponding algorithm successes rate. We use 100 different profiles

to test the trained policy of AE in the Q-learning algorithms. The total number of successful cases is recorded. The success rate is calculated as:

$$\text{Success rate} = \frac{Number\ of\ success\ profiles}{Total\ tested\ profiles} \times 100\%, \tag{39}$$

From the simulation, the success rate of QL is zero before 3000 episodes, and it is always lower than AE. When the maximum episode increases to 6000, the success rate of the QL is the same with AE, which indicates both algorithms converge to the same optimal policy if the training efforts are large enough. With the limited training episode, the proposed AE has a better training efficiency than the conventional QL.

4.3. Compare with Chance-Constrained Optimization for Conserveness

In this part, the conserveness of the proposed method and the conventional chance-constrained optimization is studied. The conserveness of the method can be regarded as the BESS capacity needed for a certain voltage regulation case. In other words, the difference between SOC usage can reflect the conserveness of the algorithms.

In Figure 9, the voltage regulation performances of chance-constrained optimization and AE are given. We use the 300 days of the load and PV data as input, and examine the proposed method in the case of a different day. The x-axis means the hour of the day, and the Y-axis is the voltage magnitude. From the simulation, both the proposed AE and chance-constrained optimization can achieve voltage regulation for the selected case.

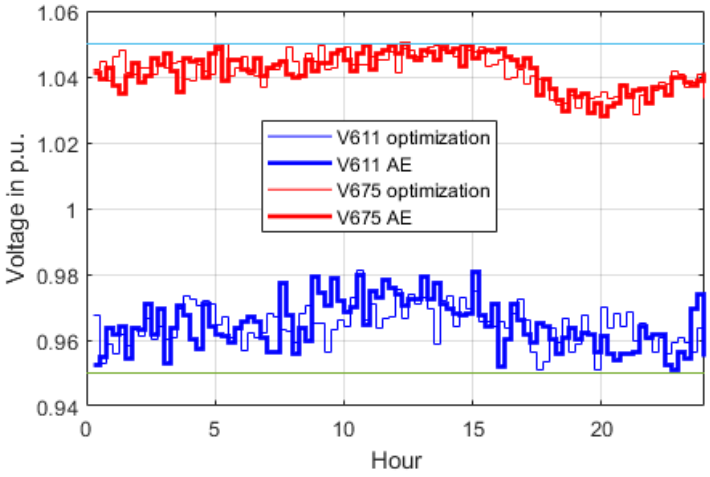


Figure 9. Rewards and average rewards of QL and AE.

The BESS usage is directly related to the SOC. High variation of the SOC profile represents the large utilization of BESS. In Figure 10, the SOC of chance-constrained optimization and AE are given. The x-axis is the hour of the day, and Y-axis is the SOC. Because the SOC is constrained in [0, 1] in this study, both the AE method and the optimization fulfill the SOC constraints. From the results, the proposed AE uses less SOC than the chance-constrained optimization in the voltage regulation, which indicates the proposed method has lower conserveness than the chance-constrained optimization. In a real application, this feature can reduce the investment of the BESSs because less capacity is required. The reason behind this conservativeness is that the chance-constrained optimization needs the BESS to operate like that for the whole training profiles. To handle all the uncertainties in the training data, the optimization solver generates a conservative operation strategy and give more stable margin on the voltage profile. This conservative characteristic improves the reliability but will cause an unnecessary BESS size.

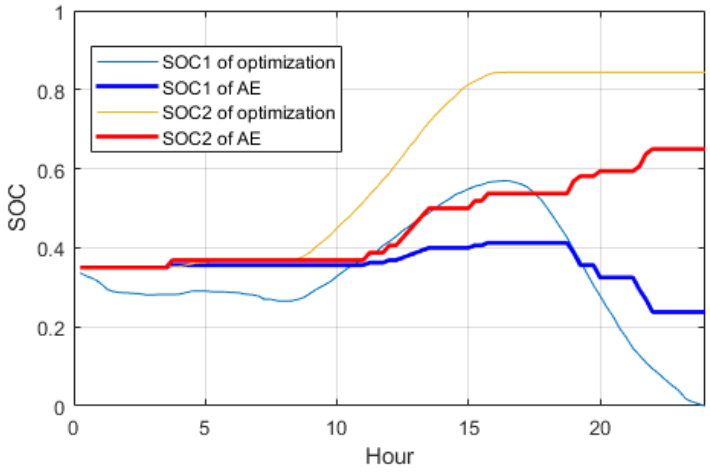


Figure 10. Conserveness reflection by SOC for AE and chance-constrained optimization.

4.4. Success Rate Comparison under Different Sizes and Weather Conditions

In this part, we examine the success rate under different BESS capacities and weather conditions. We use 300 days of the load and PV data as inputs, and exam the proposed method in the case of 150 exterior days. In the selected 150 days, the amounts of clear sky, cloudy, and overcast days are equal to 50.

Figure 11a shows the results of the success rate of the 3000 maximum training episodes, and Figure 11b illustrates the success rate of the 6000 maximum training episodes. In both figures, the proposed AE method, conventional QL and the chance-constrained optimization are compared. This comparison is to show the training efficiency.

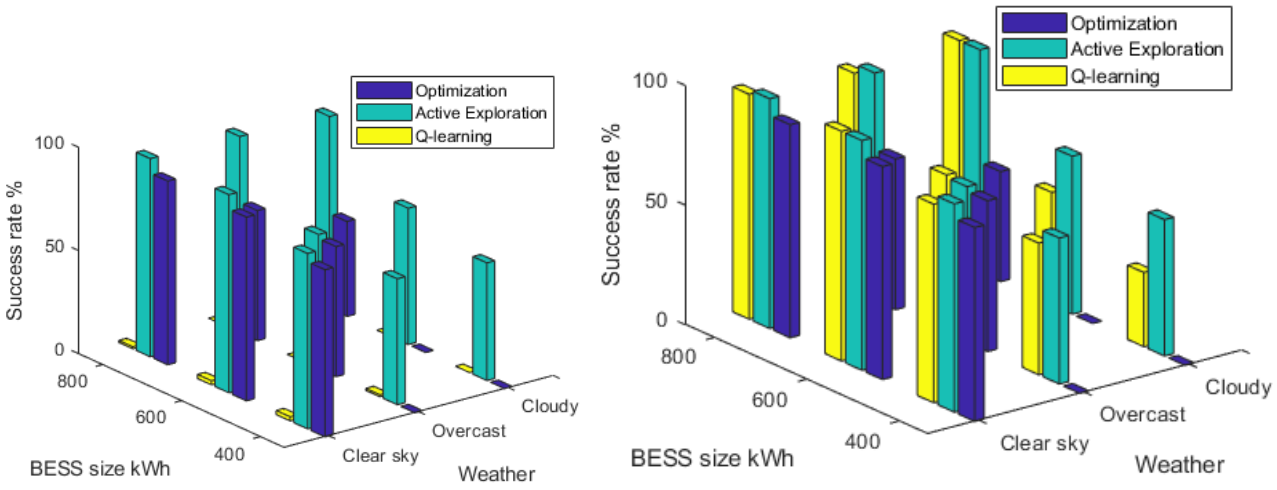


Figure 11. Success rate comparison under the different BESS size and weather conditions: (**a**) maximum episode is 3000; (**b**) maximum episode is 6000.

The success rate is the robustness metric in this work. We investigated the effect of both weather conditions and BESS size. The BESS in Section 4.1 is a 4-h BESS (200 kW with 800 kWh). In this part, a 3 h BESS and a 2 h BESS are studied. We maintained the power rate and change the energy capacity. The 3 h BESS had 200 kW and 600 kWh and the 2 h BESS had 200 kW and 400 kWh.

In Figure 11, the x-axis is the BESS size in kWh, and we have the values of 400 kWh, 600 kWh and 800 kWh. The y-axis denotes the weather condition. The z-axis represents the success rate. With a fixed weather condition, the success rate of all algorithms decreases

with the reducing BESS size. In general, the success rate of the RL is larger than the chance-constrained optimization if the training episode is large enough.

The success rate of the chance-constrained optimization is zero in some cases. The reason is that the chance-constrained optimization solver claimed these cases were infeasible because of the insufficient BESS size. The operation profiles in these cases are only based on the forecasted value without introducing uncertainties. Unfortunately, the mismatch between the forecasting data and actual net power creates voltage violations.

From Figure 11, the success rate of the chance-constrained optimization drops with weather conditions. The cloudy day has the highest forecasting mismatch, and the clear sky has the lowest mismatch. The success of the AE and Q-learning drops as well, but the speed is much lower than that of the chance-constrained optimization. That means the chance-constrained optimization is more vulnerable to uncertainties. One potential explanation could be that with the increasing uncertainty level, the difference between the training profile and the test profile grows. The chance-constrained optimization may handle cases in training profiles. However, the probability that it encounters a new worse profile during the test process increases due to the rising difference. The optimization did not meet this new worse profile before the test process, and naturally, the generated BESS operation cannot perform well. A possible solution for the chance-constrained optimization is to enlarge the number of training profiles so the optimizer can meet more uncertainties. For fairness in our comparison, the same training profiles were provided to the three algorithms.

During the optimization calculation, the chance-constrained method would like to address cases of the training profiles as much as possible. That results in over-engineered solutions (for example, a large BESS size in our application). If the size of the BESS fails to satisfy the requirement of the chance-constrained optimization, the solver would not generate any profile. However, the goal of the RL agent of Q-learning and AE is to find an average optimal policy for BESS. Although the BESS size is small, the agent can still find a policy for the operation.

When the maximum episode is set to 3000, the Q-learning cannot converge to the feasible policy, so that the corresponding success rate in Figure 11a is low. When the maximum episode becomes 6000, the success rate of Q-learning is similar to AE, which means both algorithms are converging to the optimal policy, which proves the convergence of the AE and Q-learning.

5. Conclusions

We proposed an active exploration algorithm to integrate engineering knowledge into the reinforcement learning process. The proposed algorithm can improve the training speed and accuracy compared to the traditional QL algorithm. The proposed method was used to control the BESS for voltage regulation in the IEEE 13 Node Test Feeder. The results of an empirical evaluation led to the following observations: (1) the proposed AE method can accelerate the training process and have a better success rate in voltage regulation than QL when the training time is limited. (2) The control generated by the chance-constrained optimization tends to use more battery energy in general, and it needs a larger BESS to handle uncertainties, but the proposed method uses less BESS energy and performs relatively well. (3) When the forecasting mismatch increases, the success rate of the chance-constrained optimization drops faster than the proposed method with the same BESS size. The future directions include optimal parameter tuning and involving advanced reinforcement algorithms.

Author Contributions: Z.D. conceptualized and provided resource, as well as modeling and writing of the paper; X.H. provided resource data and writing; Z.L. validated and supervised the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The main data used for this work is the bus and line parameters for the standard IEEE 13 Node Test Feeder which can be downloaded at https://cmte.ieee.org/pes-testfeeders/resources/ (accessed on 5 December 2021).

Acknowledgments: A special thanks to Ziang Zhang for his invaluable contributions to the guidance in this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hassan, A.S.; Marikkar, U.; Prabhath, G.W.; Balachandran, A.; Bandara, W.G.; Ekanayake, P.B.; Godaliyadda, R.I.; Ekanayake, J.B. A Sensitivity Matrix Approach Using Two-Stage Optimization for Voltage Regulation of LV Networks with High PV Penetration. *Energies* **2021**, *14*, 6596. [CrossRef]
2. Nour, A.M.M.; Hatata, A.Y.; Helal, A.A.; El-Saadawi, M.M. Review on voltage-violation mitigation techniques of distribution networks with distributed rooftop PV systems. *IET Gener. Transm. Distrib.* **2020**, *14*, 349–361. [CrossRef]
3. Jiang, R.; Member, S.; Wang, J.; Guan, Y.; Sets, A. Robust Unit Commitment with Wind Power and Pumped Storage Hydro. *IEEE Trans. Power Syst.* **2012**, *27*, 800–810. [CrossRef]
4. Islam, S.; Liu, P.X.; el Saddik, A. Robust control of four-rotor unmanned aerial vehicle with disturbance uncertainty. *IEEE Trans. Ind. Electron.* **2015**, *62*, 1563–1571. [CrossRef]
5. Shi, Y.; Shen, C.; Fang, H.; Li, H. Advanced control in marine mechatronic systems: A survey. *IEEE/ASME Trans. Mechatron.* **2017**, *22*, 1121–1131. [CrossRef]
6. Hong, Y.-Y.; Apolinario, G.F.D.G. Uncertainty in Unit Commitment in Power Systems: A Review of Models, Methods, and Applications. *Energies* **2021**, *14*, 6658. [CrossRef]
7. Rabiee, A.; Soroudi, A.; Keane, A. Risk-Averse Preventive Voltage Control of AC/DC Power Systems Including Wind Power Generation. *IEEE Trans. Sustain. Energy* **2015**, *6*, 1494–1505. [CrossRef]
8. Conte, F.; Massucco, S.; Saviozzi, M.; Silvestro, F. A Stochastic Optimization Method for Planning and Real-Time Control of Integrated PV-Storage Systems: Design and Experimental Validation. *IEEE Trans. Sustain. Energy* **2018**, *9*, 1188–1197. [CrossRef]
9. Li, P.; Jin, B.; Wang, D.; Zhang, B. Distribution System Voltage Control under Uncertainties Using Tractable Chance Constraints. *IEEE Trans. Power Syst.* **2019**, *34*, 5208–5216. [CrossRef]
10. Jiang, Y.; Wan, C.; Wang, J.; Song, Y.; Dong, Z.Y. Stochastic receding horizon control of active distribution networks with distributed renewables. *IEEE Trans. Power Syst.* **2019**, *34*, 1325–1341. [CrossRef]
11. Keane, A.; Ochoa, L.F.; Borges, C.L.; Ault, G.W.; Alarcon-Rodriguez, A.D.; Currie, R.A.; Pilo, F.; Dent, C.; Harrison, G.P. State-of-the-art techniques and challenges ahead for distributed generation planning and optimization. *IEEE Trans. Power Syst.* **2013**, *28*, 1493–1502. [CrossRef]
12. Xu, H.; Domínguez-García, A.D.; Sauer, P.W. Optimal Tap Setting of Voltage Regulation Transformers Using Batch Reinforcement Learning. *IEEE Trans. Power Syst.* **2020**, *35*, 1990–2001. [CrossRef]
13. Wang, W.; Yu, N.; Gao, Y.; Shi, J. Safe Off-Policy Deep Reinforcement Learning Algorithm for Volt-VAR Control in Power Distribution Systems. *IEEE Trans. Smart Grid* **2019**, *11*, 3008–3018. [CrossRef]
14. Yang, Q.; Wang, G.; Sadeghi, A.; Giannakis, G.B.; Sun, J. Two-Timescale Voltage Control in Distribution Grids Using Deep Reinforcement Learning. *IEEE Trans. Smart Grid* **2020**, *11*, 2313–2323. [CrossRef]
15. Ferreira, L.R.; Aoki, A.R.; Lambert-Torres, G. A Reinforcement Learning Approach to Solve Service Restoration and Load Management Simultaneously for Distribution Networks. *IEEE Access* **2019**, *7*, 145978–145987. [CrossRef]
16. Duan, J.; Shi, D.; Diao, R.; Li, H.; Wang, Z.; Zhang, B.; Bian, D.; Yi, Z. Deep-Reinforcement-Learning-Based Autonomous Voltage Control for Power Grid Operations. *IEEE Trans. Power Syst.* **2020**, *35*, 814–817. [CrossRef]
17. Al-Saffar, M.; Musilek, P. Reinforcement Learning-Based Distributed BESS Management for Mitigating Overvoltage Issues in Systems with High PV Penetration. *IEEE Trans. Smart Grid* **2020**, *11*, 2980–2994. [CrossRef]
18. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Moore. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
19. Zheng, Y.; Luo, S.W.; Lv, Z.A. Active exploration planning in reinforcement learning for Inverted Pendulum system control. In Proceedings of the 2006 International Conference on Machine Learning and Cybernetics, Dalian, China, 13–16 August 2006; pp. 2805–2809.
20. Främling, K. Guiding exploration by pre-existing knowledge without modifying reward. *Neural Netw.* **2007**, *20*, 736–747. [CrossRef]
21. Khamassi, M.; Velentzas, G.; Tsitsimis, T.; Tzafestas, C. Robot fast adaptation to changes in human engagement during simulated dynamic social interaction with active exploration in parameterized reinforcement learning. *IEEE Trans. Cogn. Dev. Syst.* **2018**, *10*, 881–893. [CrossRef]
22. Yu, Z.; Yang, X.; Gao, F.; Huang, J.; Tu, R.; Cui, J. A Knowledge-based reinforcement learning control approach using deep Q network for cooling tower in HVAC systems. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 1721–1726.

23. Liu, Z.; Zhang, Z. Solar forecasting by K-Nearest Neighbors method with weather classification and physical model. In Proceedings of the 2016 North American Power Symposium (NAPS), Denver, CO, USA, 18–20 September 2016; pp. 1–6.
24. Abdulla, K.; De Hoog, J.; Muenzel, V.; Suits, F.; Steer, K.; Wirth, A.; Halgamuge, S. Optimal Operation of Energy Storage Systems Considering Forecasts and Battery Degradation. *IEEE Trans. Smart Grid* **2018**, *9*, 2086–2096. [CrossRef]
25. Grant, M.; Boyd, S.; Ye, Y. CVX Users' Guide. 2009. Available online: http://www.stanford.edu/~{}boyd/software.html (accessed on 5 December 2021).
26. Sutton, R.S.; Barto, A.G. *Introduction to Reinforcement Learning*; MIT Press: Cambridge, MA, USA, 1998; Volume 135.
27. Hariri, A.; Newaz, A.; Faruque, M.O. Open-source python-OpenDSS interface for hybrid simulation of PV impact studies. *IET Gener. Transm. Distrib.* **2017**, *11*, 3125–3133. [CrossRef]