energies

MDPI

Article

Flexible Transmission Network Expansion Planning Based on DQN Algorithm

Yuhong Wang [1], Lei Chen [1], Hong Zhou [2], Xu Zhou [1], Zongsheng Zheng [1], Qi Zeng [1,*], Li Jiang [2] and Liang Lu [2]

[1] College of Electrical Engineering, Sichuan University, Chengdu 610065, China; yuhongwang@scu.edu.cn (Y.W.); chen_lei@stu.scu.edu.cn (L.C.); zhouxu@stu.scu.edu.cn (X.Z.); zongshengzheng@scu.edu.cn (Z.Z.)

[2] State Grid Southwest China Branch, Chengdu 610041, China; happypig008@sina.com (H.Z.); jianglibleach@hotmail.com (L.J.); skyluliang@126.com (L.L.)

* Correspondence: zengqi@scu.edu.cn; Tel.: +86-1778-057-3979

Abstract: Compared with static transmission network expansion planning (TNEP), multi-stage TNEP is more in line with the actual situation, but the modeling is also more complicated. This paper proposes a new multi-stage TNEP method based on the deep *Q*-network (DQN) algorithm, which can solve the multi-stage TNEP problem based on a static TNEP model. The main purpose of this research is to provide grid planners with a simple and effective multi-stage TNEP method, which is able to flexibly adjust the network expansion scheme without replanning. The proposed method takes into account the construction sequence of lines in the planning and completes the adaptive planning of lines by utilizing the interactive learning characteristics of the DQN algorithm. In order to speed up the learning efficiency of the algorithm and enable the agent to have a better judgment on the reward of the line-building action, the prioritized experience replay (PER) strategy is added to the DQN algorithm. In addition, the economy, reliability, and flexibility of the expansion scheme are considered in order to evaluate the scheme more comprehensively. The fault severity of equipment is considered on the basis of the Monte Carlo method to obtain a more comprehensive system state simulation. Finally, extensive studies are conducted with IEEE 24-bus reliability test system, and the computational results demonstrate the effectiveness and adaptability of the proposed flexible TNEP method.

Keywords: flexible transmission network expansion planning; deep Q-network; prioritized experience replay strategy; construction sequence

check for updates

Citation: Wang, Y.; Chen, L.; Zhou, H.; Zhou, X.; Zheng, Z.; Zeng, Q.; Jiang, L.; Lu, L. Flexible Transmission Network Expansion Planning Based on DQN Algorithm. *Energies* **2021**, *14*, 1944. https://doi.org/10.3390/en14071944

Academic Editor: Valentina E. Balas

Received: 9 February 2021
Accepted: 24 March 2021
Published: 1 April 2021

1. Introduction

With the rapid development of human societies, the power demand of users is also rising rapidly, and the demand for power quality is gradually rising. The continuous increase of load will change the power flow pattern of the existing power grid, which may cause potential reliability problems, such as overloads and stability issues [1]. Transmission network expansion planning (TNEP) is an effective way to solve the above problems. How to increase the transmission capacity of the transmission network, and improve the reliability and flexibility of the transmission network (as much as possible at a lower cost) is an urgent problem to be solved.

The main goal of TNEP is to expand the existing network by adding transmission lines to meet future growth in energy demand. This allows the system to maintain reliability and transmission efficiency [2]. TNEP is essentially a large-scale, non-linear, and non-convex problem. Many factors, such as alternative lines, network constraints, *N*-1 security constraints, need to be considered in the planning process. Its complexity has attracted widespread attention from scholars [3]. In 1970, the linear programming method was first introduced into the solution of TNEP [4]. Since then, a large number of scholars have carried out continuous and in-depth research on TNEP, and made good progress in planning model, planning algorithm, and other aspects.

In reference [5], the N-1 security constraints of the power grid were first considered in the planning process, and mixed integer linear programming was used to solve the problem, to improve the reliability of the transmission network with as little cost as possible. Subsequently, the uncertainties of the load and generator set are taken into consideration in the planning process. In reference [6], robust linear optimization is used to deal with uncertain factors, which further improve the reliability of the power grid. Reference [7] considered the uncertainty of wind power, established a two-stage robust planning model, and proposed a Benders' decomposition algorithm to solve it. The popularization of the Monte Carlo method opens a new chapter for the reliability analysis of TNEP. When the Monte Carlo method is applied, more reliability and safety indicators can be incorporated into the constraints and objective functions, such as expected energy not supplied (EENS) [8], security constraint unit commitment (SCUC) [9], and hierarchical reliability evaluation [10].

In essence, long-term TNEP is a type of multi-stage planning. In other words, a complex problem should be decomposed into several interrelated sub-problems, which should not only solve the problem of where to build transmission lines, but also consider when to build. Each stage should meet the requirements of economic and other indicators. The dynamic programming algorithm was proposed according to the characteristics of the multi-stage planning problem, which can effectively solve nonlinear and non-convex objective function, and deal with the change of complex constraints [11]. However, with the increases of the dimension and scale of the optimization problem, the calculation amount also increases, which is prone to the problems of curse of dimensionality and combination explosion, and difficult to deal with practical engineering problems. Therefore, scholars gradually applied the intelligent algorithms to multi-stage planning, such as teaching learning based optimization algorithm [12], high-performance hybrid genetic algorithm [13], and hybrid binary particle swarm optimization algorithm [14]. Nevertheless, existing multi-stage TNEP researches usually only consider the economic and N-1 security constraints in the evaluation of the expansion scheme. Therefore, this paper proposes a new solution to the multi-stage TNEP problems, which can comprehensively evaluate the economy, reliability, and flexibility of the expansion scheme, while ensuring convergence and low computational complexity. In addition, the proposed TNEP method can flexibly adjust the obtained scheme, which provides a lot of convenience for grid planners.

At present, a new generation of artificial intelligence technology, including machine learning, robotics and other advanced technologies, has become a research hotspot, and is profoundly affecting and changing the existing power and energy industries [15]. According to different input samples, machine learning can be divided into three categories [16]: supervised learning, unsupervised learning, and reinforcement learning (RL). Deep learning is a typical supervised learning, which can solve complex power system problems through a large amount of high-dimensional power data, such as power system transient stability assessment [17], power equipment fault diagnosis [18], and load forecasting [19,20]. However, deep learning requires a large amount of labeled data to train the neural network, which is difficult to achieve in many practical engineering problems, so it has great limitations in application. Unsupervised learning does not require data to have labels, but it is mainly used to deal with data clustering and feature learning problems [21], so it is not suitable for TNEP problems. Compared with supervised learning and unsupervised learning, RL is an active learning in essence [15]. It obtains rewards through continuous interaction with the environment, so that the agent can learn strategies to maximize rewards [22]. RL does not require labeled data. It is free to explore and develop in an unknown environment [23], with a high degree of freedom, so it can solve a variety of engineering problems. Therefore, it has become the most widely used machine learning algorithm in intelligent power systems [15]. Reference [24] combined the Q-learning algorithm with deep neural networks and proposed a deep reinforcement learning (DRL) deep Q-network (DQN) algorithm, which solved the curse of dimensionality of traditional RL algorithm in complex systems and greatly improved the learning efficiency of the agent.

At present, RL has been applied to real-time energy management of microgrid [25], smart generation control and automatic generation control [26,27], reactive power optimization for transient voltage stability [28], and other power system optimization scenarios.

Because of its high degree of freedom and weak dependence on data, DRL is very suitable for analyzing the dynamic behavior of complex systems with uncertainties. However, no scholars have applied DRL to TNEP. Compared with methods used in traditional TNEP problems such as mathematical optimization algorithms [29–31] and meta-heuristic optimization algorithms [32–34], DRL has some advantages. First, it can utilize the interactive learning characteristics of DRL to consider the construction sequence of lines and complete the adaptive planning of lines. Second, it can adjust the expansion scheme flexibly by utilizing the trained neural network without replanning, which is of great help to the grid planners and of certain guiding significance for the subsequent planning work. In addition, due to the introduction of two deep neural networks, DRL will have good convergence on the large-scale multi-stage problems.

The main contributions of this paper are listed as follows:

1. A TNEP model including the indexes of the economy, reliability, and flexibility is proposed to ensure the comprehensiveness of the scheme. Moreover, the proposed model considers *N*-1 security constraints, *N*-k faults, and the fault severity of the equipment.
2. We introduce a DQN algorithm for the first time in the solution of the TNEP problem and add prioritized experience replay (PER) strategy [35] to the traditional DQN algorithm to enhance the algorithm training effect.
3. By utilizing the interactive learning characteristics of the DQN algorithm, the construction sequence of lines is considered on the basis of a static TNEP model. In addition, it can realize the adaptive planning of the line, and flexibly adjust planned scheme according to actual need.

The rest of this paper is organized as follows: Section 2 presents the principle of the traditional DQN algorithm and introduces PER strategy. In Section 3, the objective function, constraint conditions, and indexes of the proposed TNEP model are introduced. The procedure of the proposed flexible TNEP method based on the DQN algorithm is presented in Section 4. Section 5 demonstrates the proposed method in IEEE 24-bus reliability test system and analyzes the solution process in detail. Finally, conclusions and areas for future research are given in Section 6.

2. Deep Q-Network Algorithm Based on Prioritized Experience Replay Strategy

2.1. Reinforcement Learning

The RL problem is essentially a Markov decision process (MDP), which is an interactive process in which the agent adopts random action in a deterministic environment to change its state and obtain reward. The purpose of RL is to maximize the reward with a limited number of actions to find the optimal policy. Because the decision-making process of power grid planners is similar to the MDP model, RL is suitable for solving the TNEP problems.

2.1.1. The Calculation of Value Function

Under strategy π, the agent executes action a_τ in state s_τ, and receives feedback w_τ from the environment. The feedback w_τ of the action a_τ is calculated according to the new state sτ + 1 after the action a_τ is executed. In order to reduce the influence of future rewards on the current value function, the decay factor of future rewards γ is introduced, and the value W_τ of the τ-th action is

$$W_\tau = \sum_{d=\tau}^{D} \gamma^{d-\tau} w_d \tag{1}$$

The action's Q value can be calculated based on its W_τ. The state–action value function $Q_\pi(s,a)$ represents the expected return value generated by executing action a in the current state s under the strategy π

$$Q_\pi(s,a) = E_\pi[W_\tau | S_\tau = s, A_\tau = a] \tag{2}$$

It can be seen that the value function is calculated in a recursive manner, which also shows that RL algorithm is suitable for multi-stage planning problems. If the expected return value of a strategy in all states is not inferior to other strategies, it is called the optimal strategy. There may be more than one optimal strategy. π^* is used to represent the set of optimal strategies. They share the same optimal state–state value function $Q^*(s,a)$, which is the value function with the largest value among all strategies. The expression is given as follows

$$Q^*(s,a) = \max_{\pi^*} Q_{\pi^*}(s,a) \tag{3}$$

The Bellman equation (BE) of the optimal strategy can be obtained

$$Q^*(s_\tau, a_\tau) = E_\pi[w_{\tau+1} + \gamma \max_{\pi^*} Q^*(s_{\tau+1}, a_{\tau+1}) | s_\tau, a_\tau] \tag{4}$$

2.1.1.1. *ε*-Greedy Action Selection Strategy

In the iterative process of RL, the state–action value function $Q(s,a)$ representing the value of the action a_τ selected under the state s_τ will be updated in real time. In order to enhance the global search ability of the algorithm, this paper adopts the *ε*-greedy action selection strategy $\pi(s)$

$$\pi(s) = \begin{cases} \mathrm{argmax}_a Q(s,a) & 0 \le \mu < \varepsilon \\ \forall a \in A & \varepsilon \le \mu \le 1 \end{cases} \tag{5}$$

When $\mu < \varepsilon$, the action a with the largest Q value is selected, otherwise, it is a random action. The update of Q-table based on temporal difference (TD) prediction (calculate the Q values based on the current state s_τ and the next state $s_{\tau\,+1}$) is

$$Q(s_\tau, a_\tau) \leftarrow Q(s_\tau, a_\tau) + \alpha\left[w + \gamma \max_{a_{\tau+1}} Q(s_{\tau+1}, a_{\tau+1}) - Q(s_\tau, a_\tau)\right] \tag{6}$$

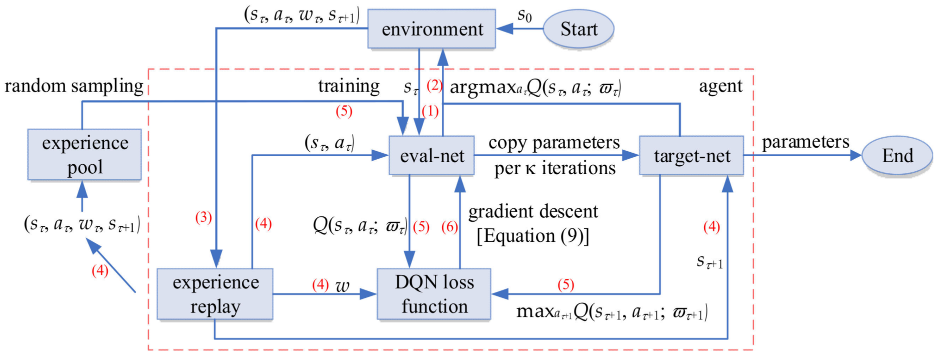
where $s_{\tau+1}$ represents the new state after selecting action a_τ in state s_τ; $a_{\tau+1}$ represents the most valuable action in state $s_{\tau+1}$. $w + \gamma \max_{a\tau_{+1}} Q(s_{\tau+1},a_{\tau+1})$ represents the actual value of Q, and $Q(s_\tau,a_\tau)$ represents the estimated value of Q. The difference between the absolute values of the actual value of Q and the estimated value of Q is called TD error Δ_τ. The smaller the Δ_τ, the better the training effect of the agent.

2.2. *Deep Q Network Algorithm*

The Q-learning algorithm of traditional RL is difficult to solve the problems of large-scale MDP or continuous space MDP due to Q-table's curse of dimensionality of complex networks. For this reason, the DeepMind proposed DQN algorithm to approximate the Q-Table [24]. Moreover, the Q value of each action can be predicted by only inputting the current state s_τ and calling Q-table once. The DQN algorithm combines RL with neural network to fit the value function $Q(s,a)$ through the deep neural network $Q(s,a;\varpi)$, where ϖ is the weight of the neural network.

In the DQN algorithm, the agent is the part responsible for learning, and the environment is the part where the agent interacts with specific problems. The main function of the DQN algorithm is to make the agent learn the best possible action and make the subsequent rewards as large as possible. The function of the agent is to complete the selection of action a_τ and the training of the neural network. The function of the environment is to complete the update of state s_τ and the calculation of reward w_τ. The DQN algorithm will generate two multilayer perceptron neural networks with the same structure, eval-net and target-net, which are used to calculate the actual value of Q and the estimated value of Q respectively. The agent selects the next action based on these two neural networks. The agent's experience $(s_\tau,a_\tau,w_\tau,s_{\tau\,+1})$ is stored in the experience pool and randomly sampled for training eval-net. Furthermore, the eval-net's parameters are continuously updated based on the loss function, and the target-net's parameters are copied from eval-net per κ iterations (one iteration includes action selection, reward calculation, network structure update, and eval-net update), thus guaranteeing the convergence of the DQN algorithm.

The specific process is shown in Figure 1. The sequence of each step in an iteration has been marked with red serial numbers.



$(s_\tau, a_\tau, w_\tau, s_{\tau+1})$
environment
s_0
Start
random sampling
training
s_τ
(2)
$\mathrm{argmax}_{a_\tau} Q(s_\tau, a_\tau; \varpi_\tau)$
agent
(5)
(1)
experience pool
(s_τ, a_τ)
eval-net
copy parameters per κ iterations
target-net
parameters
End
$(s_\tau, a_\tau, w_\tau, s_{\tau+1})$
(3)
(4)
$Q(s_\tau, a_\tau; \varpi_\tau)$
(5)
(6)
gradient descent [Equation (9)]
(4)
$s_{\tau+1}$
(4)
experience replay
(4) w
DQN loss function
(5)
$\mathrm{max}_{a_{\tau+1}} Q(s_{\tau+1}, a_{\tau+1}; \varpi_{\tau+1})$

Figure 1. Deep *Q*-network (DQN) algorithm flow chart.

In the DQN algorithm, the calculation of Q network label $Q_{\max}$ is

$$Q_{\max} = \begin{cases} w_\tau, & \text{if episode terminates at next step} \\ w_\tau + \gamma \max_{a_{\tau+1}} Q_{\text{target}}(s_{\tau+1}, a_{\tau+1}), & \text{otherwise} \end{cases} \tag{7}$$

the update formula of value function is

$$Q_{\text{eval}}(s_\tau, a_\tau) \leftarrow Q_{\text{eval}}(s_\tau, a_\tau) + \alpha[Q_{\max} - Q_{\text{eval}}(s_\tau, a_\tau)] \tag{8}$$

and the update formula of the eval-net's weight ϖ is

$$\varpi_{\tau+1} = \varpi_\tau + \alpha[Q_{\max} - Q_{\text{eval}}(s_\tau, a_\tau; \varpi)]\nabla_\varpi Q_{\text{eval}}(s_\tau, a_\tau; \varpi) \tag{9}$$

In the training process of the eval-net, the loss function uses the mean square error function

$$L(\varpi) = E[Q_{\max} - Q_{\text{eval}}(s_\tau, a_\tau; \varpi)]^2 \tag{10}$$

In the traditional DQN algorithm, each experience replay is a random extraction action. However, different experience has different training effect on the neural network, and the training effect of experience with more extreme Δ_τ will be better. Therefore, DeepMind proposed PER strategy [35], which sorts the experience in the experience pool according to the importance of Δ_τ. The experience closer to the head and end is more important and will have a higher priority in sampling. In this way, more learning-worthy experience can be effectively extracted, to improve the learning efficiency of the agent. This paper prioritizes experience based on Δ_τ, and defines the priority of experience τ as

$$p_\tau = 1/\text{rank}(|\Delta_\tau|) \tag{11}$$

Moreover, in order to avoid the loss of diversity and over-fitting due to frequently repeated sampling of the experience with the front serial number, this paper combines PER strategy with stochastic sampling to ensure that the experience with low priority can also be selected. The probability of extracting experience τ is

$$p(\tau) = p_\tau^\varphi \Big/ \sum_{\tau=1}^{\Gamma} p_\tau^\varphi \tag{12}$$

When φ = 0, it is random sampling; Γ represents the size of the playback experience pool.

3. Transmission Network Expansion Planning Model

This section mainly introduces the calculation of the comprehensive equivalent cost $f(s_\tau)$ of the TNEP model. In the TNEP planning method based on the DQN algorithm, the reward w_τ of each line-building or line-deleting action a_τ is calculated from the $f(s_{\tau+1})$ of the new network structure $s_{\tau+1}$ after the action a_τ is executed (the specific formula is Equation (35) in Section 4), and then the Q value of the action a_τ is calculated according to Equations (7) and (8). The calculation of the scheme's $f(s_\tau)$ is part of the environment in the DQN algorithm. In addition, the proposed TNEP model is a static planning model rather than a multi-stage planning model. The consideration of the construction sequence of lines is realized by utilizing the interactive learning characteristics of the DQN algorithm. Since the DQN algorithm considers the influence of subsequent actions when calculating the Q value, the influence of subsequent actions on the overall expansion scheme can be considered in each action selection when utilizing the DQN algorithm to solve the multi-stage TNEP problem. The effect of multi-stage planning will not be affected by the simplicity of the planning model.

3.1. Objective Function

The evaluation of the TNEP scheme in this paper comprehensively considers the economy, reliability, and flexibility of the system. The economic indexes mainly include the annual equivalent line construction investment cost C_{in}, the operation and maintenance cost C_o, and the network loss cost C_{loss}. The reliability index EENS is transformed into the reliability cost C_{EENS}, and the flexibility of the planning scheme is evaluated by the flexibility index average normalized risk index of bus voltages (ANRIV) ζ_{ANRIV}. The comprehensive equivalent cost $f(s_\tau)$ is obtained by combining the above indexes, and it is used as the objective function for TNEP

$$\min f(s_\tau) = (C_{in} + C_o + C_{\text{loss}} + C_{\text{EENS}})(1 + \zeta_{\text{ANRIV}}) \tag{13}$$

$$\mathrm{C}_{in} = \lambda_{in} \sum\nolimits_{\mathrm{l} \in \mathrm{S}} \beta_{\mathrm{l}} c_{\mathrm{l}} L_{\mathrm{l}} \tag{14}$$

$$\lambda_{in} = \frac{\xi(1+\xi)^{y_0}}{(1+\xi)^{y_0} - 1}(1+\xi)^{y_1} \tag{15}$$

$$C_o = \lambda_o \sum\nolimits_{l \in \Psi} L_l + K_{\mathrm{G}} T \sum\nolimits_{i=1}^{N} P_{\mathrm{G},i} \tag{16}$$

$$C_{\text{loss}} = K_{\text{loss}} T \sum\nolimits_{h}^{H} \sum\nolimits_{i=1}^{N} \sum\nolimits_{j \in c(i)} I_{ij,h}^2 r_{ij,h} \tag{17}$$

$$C_{\text{EENS}} = K_{\text{EENS}} P_{\text{EENS}} \tag{18}$$

3.2. Constraints

When the system operates normally, its power flow constraints, generator output constraints, line operating state constraints, and bus voltage constraints can be formulated as follows:

$$P_{\mathrm{G},i} - P_{\text{load},i} - \sum\nolimits_{h}^{H} \sum\nolimits_{j \in c(i)} b_{ij,h} \theta_j = 0 \tag{19}$$

$$P_{\mathrm{G},i}^{\min} \leq P_{\mathrm{G},i} \leq P_{\mathrm{G},i}^{\max} \tag{20}$$

$$\left| P_{ij,h} \right| \leq P_{ij,h}^{\max} \tag{21}$$

$$U_i^{\min} \leq U_i \leq U_i^{\max} \tag{22}$$

3.3. *N*-k Fault Power Flow Calculation Model

The load shedding is taken as the objective function to obtain the optimal power flow of the system under *N*-k fault. The *N*-k fault power flow calculation model considering power flow constraints, generator active power output constraints, load shedding constraints, and line power flow constraints can be formulated as follows:

$$\min f_z = \sum\nolimits_{i \in N} P_{i,z} \tag{23}$$

$$\text{s.t.} \begin{cases} P_{\mathrm{G},i,z} - P_{\mathrm{load},i} + P_{i,z} - \sum_{h}^{H} \sum_{j \in c(i)} b_{ij,h} \theta_{j,z} = 0 \\ 0 \le P_{\mathrm{G},i,z} \le P_{\mathrm{G},i}^{\max} \\ 0 \le P_{i,z} \le P_{\mathrm{load},i} \\ \left| P_{ij,h,z} \right| \le P_{ij,h}^{\max} \end{cases} \tag{24}$$

3.4. EENS Cost Considering the Fault Severity

Based on the Monte Carlo method, this paper considers the fault severity of equipment, and improves the average and scattered sampling method to obtain faster sampling efficiency and more comprehensive system state simulation.

A uniformly distributed random number μ in the interval [0, 1] is generated to simulate the operating state of a certain equipment e. If the equipment e is a transmission line, the operating state δ_e of the equipment can be expressed as

$$\delta_e = \begin{cases} 0, & 0 \le \mu \le p_{f,e} \\ 1, & p_{f,e} < \mu \le 1 \end{cases} \tag{25}$$

If the equipment e is a generator set, in addition to the operation of the whole unit and the shutdown of the whole unit, there may also be some states of shutdown of some units. Divide the interval [0, 1] into y sub-intervals of equal length to simulate the different operating states of the unit. The operating state δ_e and active power output of the unit can be expressed as follows:

$$\delta_e = \begin{cases} 0, & others \\ 1, & 0 \le \mu \le P_{\mathrm{f,e}}/y \\ 2, & 1/y \le \mu \le (1 + P_{\mathrm{f,e}})/y \\ \vdots & \\ y, & (y-1)/y \le \mu \le (y - 1 + P_{\mathrm{f,e}})/y \end{cases} \tag{26}$$

and the active power output of generator set e under state z is

$$P_{\mathrm{G},e,z} = (1 - \delta_e / y) P_{\mathrm{G},e} \tag{27}$$

The operation state M_z of the transmission network can be obtained by sampling all of the equipment above

$$M_z = \{\delta_1, \delta_2, \cdots, \delta_F\} \tag{28}$$

After enough sampling, the occurrence frequency of state M_z can be taken as an unbiased estimate of its occurrence probability $p(z)$

$$p(z) = n(M_z)/n_{\mathrm{total}} \tag{29}$$

Therefore, the total load shedding P_{EENS} of the transmission network is

$$P_{\mathrm{EENS}} = T \sum\nolimits_{z \in \Phi} \left(p(z) \sum\nolimits_{i \in N} P_{i,z} \right) \tag{30}$$

Combined with Equation (30), Equation (18) can be transformed into

$$C_{\mathrm{EENS}} = K_{\mathrm{EENS}} T \sum\nolimits_{z \in \Phi} \left(p(z) \sum\nolimits_{i \in N} P_{i,z} \right) \tag{31}$$

3.5. The Average Normalized Risk Index of Bus Voltages

When a fault occurs in the system, if the bus voltage is higher than the rated value U_i^{rate}, it is considered as risk-free, as shown in Figure 2. The difference from [36] is that when the bus voltage is lower than $U_i^{\min}$, the interval normalized risk index (NRI) value increases exponentially, which can increase the influence of unstable voltage on the NRI. The NRI $\zeta_{\text{ V},i}$ of the voltage at bus i is

$$\zeta_{V,i} = \begin{cases} 0, & U_i \geq U_i^{\text{rate}} \\ \frac{U_i^{\text{rate}} - U_i}{U_i^{\text{rate}} - U_i^{\min}}, & U_i^{\min} \leq U_i \leq U_i^{\text{rate}} \\ \exp\left(\frac{U_i^{\min} - U_i}{U_i^{\text{rate}} - U_i^{\min}}\right), & U_i < U_i^{\min} \end{cases} \tag{32}$$

When equipment e is outage, the NRI at bus i is

$$\zeta_{\text{V},i,e} = p_{f,e}\zeta_{\text{V},i} \tag{33}$$

To evaluate the overall flexibility of the TNEP scheme, the flexibility index ζ_{ANRIV} is obtained by calculating the system power flow under all N-1 faults

$$\zeta_{\text{ANRIV}} = \eta_{\text{ANRIV}}\left(\sum_{e\in F}\sum_{i\in N}\zeta_{\text{V},i,e}\right)/(F\cdot N) \tag{34}$$

The smaller the ζ_{ANRIV} of the expansion scheme, the stronger the adaptability to equipment outages, and the better its flexibility.

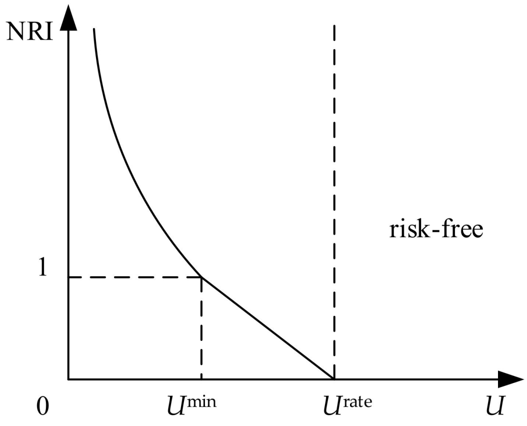


Figure 2. Normalized risk index (NRI) of bus voltage.

4. Flexible TNEP Based on DQN Algorithm

4.1. Algorithm Architecture Design

The TNEP framework based on the DQN algorithm is shown in Figure 3. In fact, the two neural networks are also parts of the agent. However, in order to make readers better understand the process of using the DQN algorithm in TNEP, we put the two neural networks outside the agent. The network structure s_τ consists of the construction state β_l of the buildable lines; the action set is the set S of the buildable lines. According to Equations (5), (7) and (8), the agent selects the line-building or line-deleting action a_τ according to the existing network structure s_τ and the Q values of actions, and then the action a_τ is fed back to the planning environment of the transmission network. The environment performs N-1 analysis on the new network structure $\text{s}_{\tau+1}$, calculates various costs and indexes, obtains comprehensive equivalent cost $f(s_\tau)$, and calculates action reward w_τ. The environment feeds back the new network structure $\text{s}_{\tau+1}$ and action reward w_τ to the agent, who collects experience. The selection of the line-building or line-deleting action a_τ, the update of the network state s_τ, and the calculation of action reward w_τ, together form the MDP of TNEP. After collecting a certain amount of experience, it will extract experience training eval-net according to PER strategy. The agent learns and gives an optimal action plan, and copies the eval-net's parameters to target-net per κ iterations.

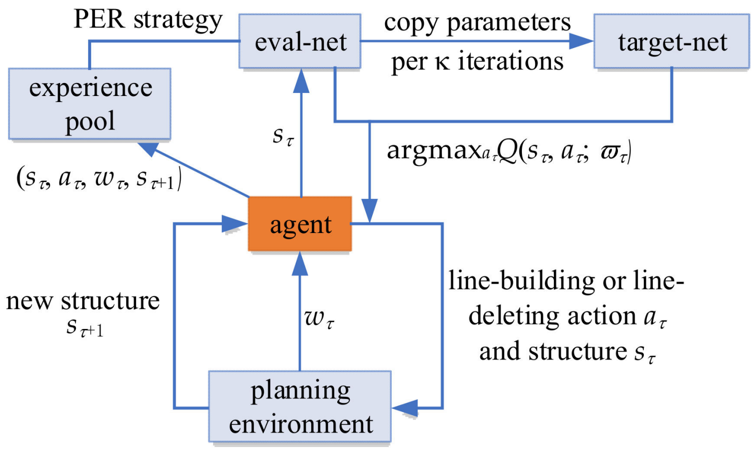

PER strategy
eval-net
copy parameters
per κ iterations
target-net
experience pool
s_τ
$\text{argmax}_{a_\tau} Q(s_\tau, a_\tau; \varpi_\tau)$
$(s_\tau, a_\tau, w_\tau, s_{\tau+1})$
agent
new structure
$s_{\tau+1}$
w_τ
line-building or line-deleting action a_τ and structure s_τ
planning environment

Figure 3. Transmission network expansion planning (TNEP) framework based on the DQN algorithm.

The correct setting of the reward function is crucial for the DQN algorithm. The reward w_τ of line-building or line-deleting action a_τ is calculated based on the environment's evaluation of the new structure $s_{\tau+1}$. This paper judges the quality of the network structure according to the magnitude of $f(s_\tau)$. Therefore, this paper first sets a larger benchmark cost to perform iterative learning to obtain the comprehensive equivalent cost of a suitable scheme. The final benchmark cost f_{base} is appropriately increased on this basis, and the N-1 security constraints is taken into account, so that the reward w_τ of action a_τ is

$$w_\tau = \begin{cases} f_{\text{base}} - f(s_{\tau+1}), & \text{satisfy } N-1 \text{ security constraints} \\ -100, & \text{otherwise} \end{cases} \tag{35}$$

Therefore, the expansion scheme that not satisfies N-1 security constraints will result in a large negative reward for the last action. Moreover, the scheme whose comprehensive equivalent cost is lower than the benchmark cost will make the last action get a positive reward, otherwise the last action will get a negative reward. Therefore, the agent will explore in the direction where the expansion scheme satisfies N-1 security constraints and $f(s_\tau)$ is smaller.

4.2. Planning Process

The interaction mechanism between the planning environment and the agent has been introduced before. The main purpose of this paper is to study a flexible TNEP method, so the treatment of multi-stage planning is relatively brief. Whenever an action is selected, the network structure will be updated, $f(s_\tau)$ and reward w_τ will be calculated. However, this paper only considers the sequence of line-building actions, not the precise time of each line's construction. In addition, each scheme is calculated as a static programming. The detailed planning procedure of the proposed flexible TNEP method based on the DQN algorithm is provided in Figure 4. In each iteration, only one line's construction state will be changed, which ensures the convergence of the DQN algorithm. Whenever an expansion scheme $w_\tau > 0$, an episode (iteration round) ends. In addition, this paper creates a database to save all the expansion schemes that satisfy N-1 security constraints and calls them during the planning process. Therefore, repeated schemes will not be recalculated, saving a lot of time.

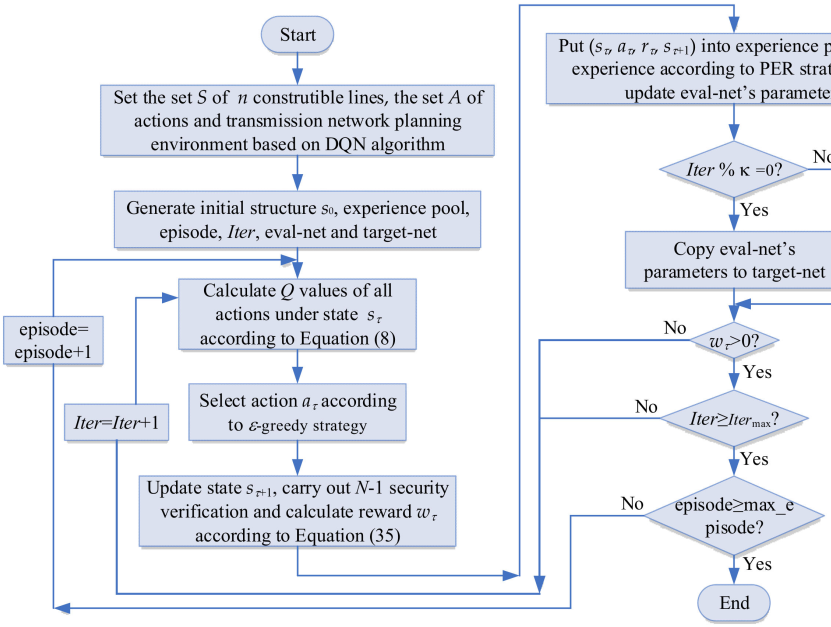

Start
Set the set S of n construtible lines, the set A of actions and transmission network planning environment based on DQN algorithm
Generate initial structure s_0, experience pool, episode, Iter, eval-net and target-net
Calculate Q values of all actions under state s_τ according to Equation (8)
Select action a_τ according to ε-greedy strategy
Update state $s_{\tau+1}$, carry out N-1 security verification and calculate reward w_τ according to Equation (35)
episode= episode+1
Iter=Iter+1
Put (s_τ, a_τ, r_τ, $s_{\tau+1}$) into experience pool, select experience according to PER strategy, and update eval-net's parameters
Iter % κ =0?
No
Yes
Copy eval-net's parameters to target-net
No
w_τ>0?
Yes
No
Iter≥$Iter_{max}$?
Yes
No
episode≥max_e pisode?
Yes
End

Figure 4. Planning procedure of the proposed flexible TNEP method.

5. Case Study

In this paper, the IEEE 24-bus reliability test system [37] is selected for calculation and analysis. The system consists of 38 power transmission lines and 32 generator sets, of which 17 buses carry loads, with a total load value of 2850 MW. The system can be divided into 230 kV in the north area and 138 kV in the south area, connected by five transformer branches. The to-be-selected set of buildable lines in this paper includes 38 original transmission lines of the system and 50 to-be-selected lines. The parameters of the 38 original transmission lines of the system can be found in [37], and these lines are numbered from 1 to 38. Some parameters of the 50 to-be-selected lines are shown in Appendix A, where the investment cost has been converted to an equivalent annual cost.

In order to verify the applicability and advantages of the DQN algorithm in TNEP, this paper designs three experiments. Experiment 1 is the TNEP on the original network, experiment 2 is the modification of the expansion scheme, and experiment 3 is the subsequent line planning. In the three experiments, the power generations and loads are increased by 1.5 times, so it is necessary to increase the transmission lines to ensure the safe and stable operation of the transmission network.

5.1. Experiment 1

In order to verify the performance improvement of the algorithm brought by PER strategy, the DQN algorithm based on random sampling and PER strategy are respectively used to conduct experiment 1. The algorithm parameter settings are the same, the maximum iteration number of one episode is $Iter_{\max}$ = 200, the maximum iteration round is max_episode = 200, the annual maximum load utilization hours is T = 5000 h, the annual value coefficient of line operation and maintenance cost is λ_o = 0.05, the power generation cost is K_{G} = 20 $/MWh, the network power loss price is K_{loss}=40 $/MWh, the load shedding cost is K_{EENS} = 100 $/MWh, the number of equal scattered sampling sections is y = 4, and the number of Monte Carlo sampling is 2000. Due to the long calculation time of Monte Carlo sampling, in order to save program running time and ensure the quality of the expansion scheme, for the scheme of $\zeta_{\text{ANRIV}} > 1$, directly set C_{EENS} = 40 M$, and the agent will get negative reward.

Figure 5 shows the comparison of the total number of iterations in the first 50 episodes before and after the introduction of PER strategy. If no feasible scheme is found in the learning of one episode, the number of this episode's iterations will be equal to $Iter_{\max}$. It can be seen from Figure 5 that after PER strategy is introduced, the number of iterations without a feasible scheme is reduced, and the total number of iterations in the first 50 episodes is reduced by 30%. All these verify the help of PER strategy to improve the learning efficiency of the agent. Figure 6 shows the comparison of each line-building action's Q value under the two experience extraction strategies in the last episode.

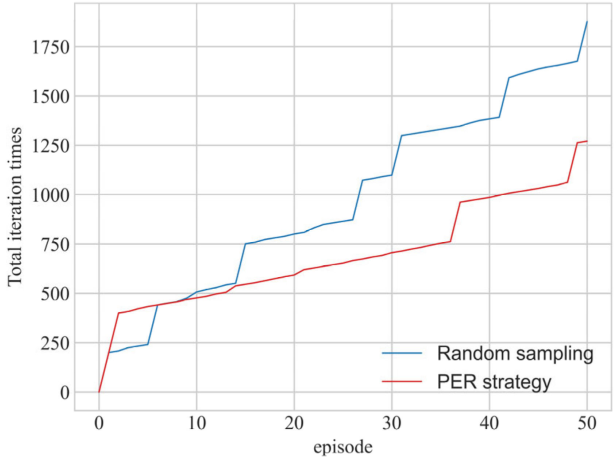


Figure 5. Comparison of the number of iterations before and after the introduction of prioritized experience replay (PER) strategy.

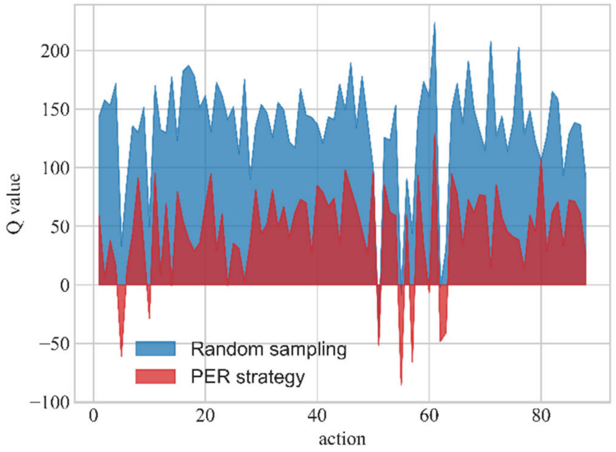


Figure 6. Comparison of the Q values under the two experience extraction strategies in the last episode.

The magnitude of the action's Q value reflects its improvement to the transmission network system. The calculation of the Q value in the eval-net of the DQN algorithm is given in Equations (7) and (8). Since there is a certain error in predicting $Q_{\max}$ by the neural network and the calculation of $Q_{\max}$ may base on the Q_{target} with the largest error, the Q value calculation of the DQN algorithm will have an overestimation problem. Figure 6 shows that after PER strategy is introduced, most of the Q values of the same line-building actions on the initial network are decreasing, indicating that the situation of overestimation has been improved. In addition, under the two experience extraction strategies, the agents have basically the same judgments on the line construction, and the Q values have been dropped by 30% on average after the introduction of PER strategy. By observing the maximum Q values under the two experience extraction strategies in Figure 6, it can be found that the two maximum Q values are both the 61st line-building action, so the first line to be built on the initial network is the 61st line. According to Appendix A, it can be seen that this line is the transmission line between bus 6 and 7 (line 6–7). Moreover, bus 7 is connected to other buses with only one line, which cannot satisfy N-1 security constraints. Therefore, adding a transmission line connected to bus 7 can effectively reduce the reliability cost and ζ_{ANRIV}, and the Q value of such a line-building action will be relatively larger. When PER strategy is introduced, each expansion scheme that satisfies N-1 security constraints and its indexes values are recorded in sequence. The changes of each index with iterations are presented in Figure 7. The darker the color, the more data nearby. The subgraphs on the right are also used to reflect the distribution of data.

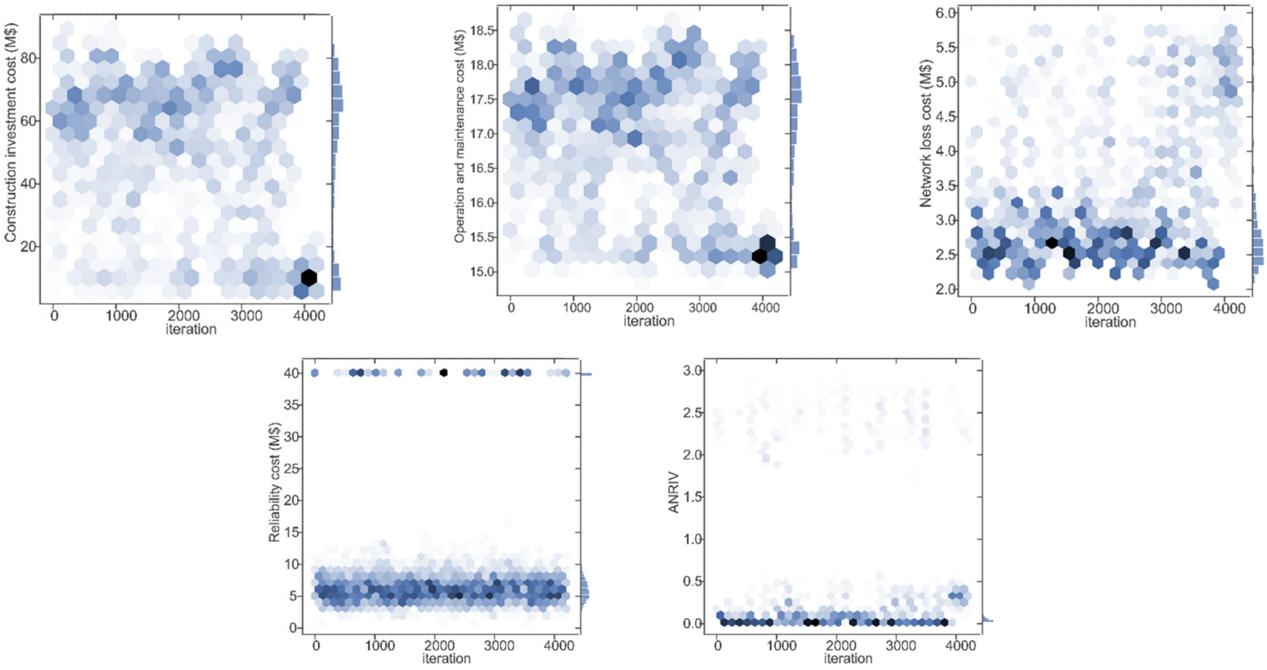


Figure 7. Changes of each index with iterations.

The first two pictures in Figure 7 show that in the first 3500 iterations of the algorithm, many expansion schemes with high construction investment costs and operation and maintenance costs were recorded. This is because at the beginning of the algorithm, the agent cannot clearly determine which line-building actions can effectively reduce reliability costs and ζ_{ANRIV}, and will build more lines to meet the requirements of reliability and flexibility. In addition, due to the construction of more lines, the power flow in the transmission network will be more balanced, and the network loss cost will be lower. The data distribution in the third subgraph in Figure 7 also verifies this result. High construction investment cost leads to the high $f(s_T)$ of the scheme, so no feasible solution can be found in such episodes and the database will record a large number of infeasible schemes at the beginning of the algorithm. As the training continues, the agent can gradually judge the quality of each line-building action, so that it can find feasible schemes faster, and the distribution of the index values has also stabilized, which also verifies the powerful learning effect of agent in the DQN algorithm based on PER strategy.

Since the network cannot satisfies N-1 security constraints without expansion planning, the reliability cost is 12.59 M$, and the flexibility index is $\zeta_{\mathrm{ANRIV}} = 1.57$, its reliability and flexibility are very poor. Therefore, the original network structure needs to be expanded to enhance system reliability and flexibility. In order to compare with the traditional planning algorithm, this paper carries out planning under two different planning scenarios. The planning schemes in this paper are sorted according to the sequence of line construction and compared with reference [38] in Table 1. The scheme in [38] is obtained by a mixed-integer linear programming approach. The line investment cost and the operation cost of generation units are taken as the objective function in the planning in [38], and only N-1 security constraints are taken into account.

Table 1. IEEE 24-bus reliability test system planning scheme comparison.

	Planning Scheme	C_{in}/M$	C_o/M$	C_{loss}/M$	C_{EENS}/M$	ζ_{ANRIV}	$f(s_\tau)$/M$
Scheme 1	6–7,19–21,19–22,7–8, 2–9,22–23,6–10	6.73	15.02	4.66	4.27	0.21	37.22
Scheme 2	6–7,19–21,19–22,18–21, 3–6,1–5,20–23	8.99	15.17	4.83	1.95	0.40	43.35
Reference [38]	3–24,7–8,14–16,15–21, 15–24,16–17,20–23	10.44	15.20	4.62	4.14	0.72	59.18

In Table 1, scheme 1 is the scheme with the least comprehensive equivalent cost. Since the flexibility index ζ_{ANRIV} is not much different among the schemes with lower comprehensive equivalent costs, this paper chooses the scheme with the least reliability cost as scheme 2. Both of these schemes are obtained by the DQN algorithm based on PER strategy, and only the objective functions of the planning are different. The comparison between the scheme in [38] and scheme 1 selected in this paper is shown in Figure 8.

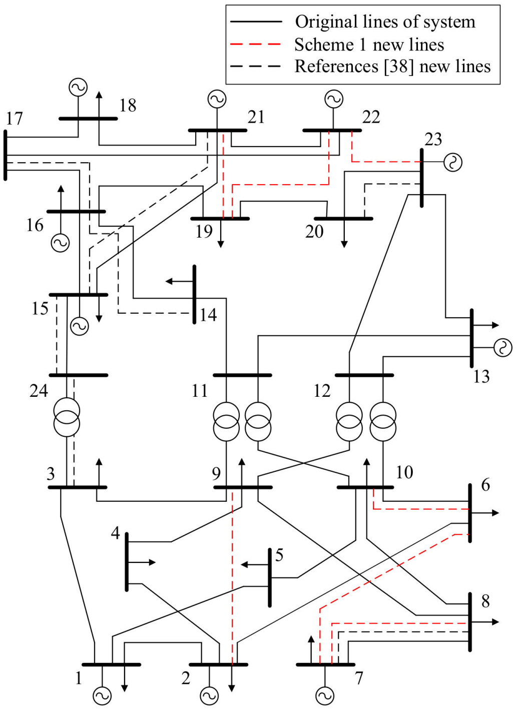


Figure 8. Comparison of planning schemes.

It can be seen from Figure 8 that the two schemes have only one identical extension line, which is line 7–8. The scheme in [38] mainly focuses on the power exchange within the northern area and between the northern and southern areas. The south area only adds line 7–8 to ensure that the system can satisfies N-1 security constraints. Scheme 1 is mainly to strengthen the power exchange within the northern and southern areas, and does not enhance the power exchange between the northern and southern areas. Three transmission lines have been added between the buses on the right side of the northern area, and several transmission lines have been added on the right side of the southern area and between buses 2–9. Since the network loss and N-1 security constraints are more important in the planning in [38], the network loss cost and reliability cost of the scheme in [38] are relatively lower. However, compared with the two planning schemes in this paper, its low bus voltage after the fault occurs will be more serious, so there is still some space for optimization.

The first three line-building actions of scheme 1 and scheme 2 are the same, but due to the different number of episodes, the maximum value judgment will be different. There are differences in subsequent line construction, but the line-building strategies are the same. From the line construction sequence of the two schemes, the agent chooses to add a line connected to bus 7 first to solve the problem that *N*-1 security constraints are not satisfied at bus 7, thereby greatly enhancing the system's reliability and flexibility. Subsequently, the agent chooses to add two transmission lines in the northern area to enhance the reliability of the northern area. Compared with scheme 1, although scheme 2 has a 2.32 M$ lower reliability cost, other indexes are larger. The ζ_{ANRIV} of scheme 2 performs poorly, so the overall system bus voltage will be lower when a fault occurs, and the overall economy will be worse.

In order to verify the improvement of power flow distribution after faults in scheme 1, in this paper we carried out multiple sets of *N*-1 security constraints. Figure 9 shows the comparison of network loss after different lines are cutting. It can be seen that the network loss of scheme 1 is generally smaller after the line faults. Therefore, the power flow distribution of scheme 1's structure is more reasonable and reliable, and the space for power flow adjustment according to the demand is larger, which all prove that scheme 1 has better reliability and flexibility.

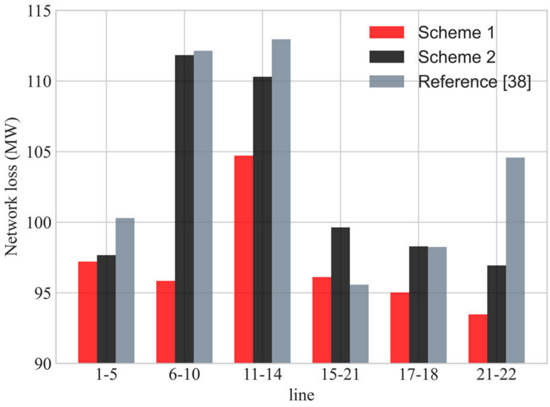


Figure 9. Comparison of network loss after cutting different lines.

Compared with mathematical optimization algorithms and meta-heuristic optimization algorithms, the planning method proposed in this paper can visualize the data generated during the planning process. Due to the coordinated calculation of two deep neural networks with the same structure, the DQN algorithm can also have strong convergence in large-scale planning problems. In addition, the use of ε-greedy action selection strategy makes the DQN algorithm have a large search space.

5.2. Experiment 2

In actual engineering, the expansion scheme may need to be adjusted during the construction period due to various reasons. For example, during the implementation of scheme 1, line 7–8 that should have been fourth added could not be built due to some special reasons. At this time, the remaining expansion scheme needs to be re-planned. Conventional planning methods need to make certain modifications to the model and then plan again. The TNEP method based on DRL in this paper does not need to be re-planned. It only needs to import the trained neural network parameters and input the extended network structure to obtain the *Q* value of each subsequent line-building or line-deleting action. So that planners can select one line-building action with high *Q* value combined with a variety of factors, which has strong flexibility.

Inputting the network structure of the first three lines in the scheme 1 obtained in experiment 1 (lines 6–7, 19–21 and 19–22) into the neural network trained by the DQN algorithm, the *Q* value of each action is obtained as shown in Figure 10. It can be seen from Figure 10 that the most appropriate line to build is the 11th line (line 7–8), which is consistent with the line selected in scheme 1. The Q value of the 61st line (line 6–7) is very small, because this line is the first line built in scheme 1. Choosing this action means deleting this line, which will make the network not satisfy N-1 security constraints and get a negative reward. Therefore, the Q value of this action is very small. To verify the correctness of the action selection, line 11–14 is selected from other actions with high Q values, and line 17–18 is selected from actions with low Q values. The indexes and costs after the construction of line 7–8, and these two lines, are shown in Table 2.

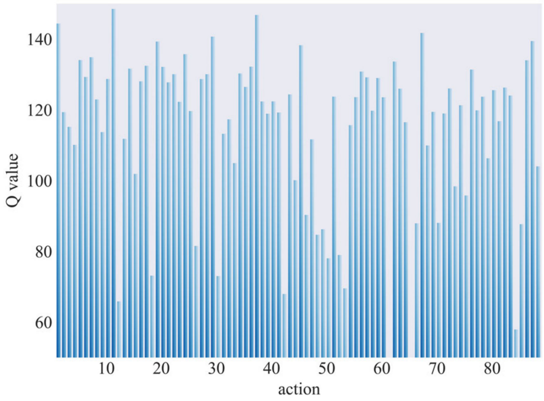


Figure 10. Q value of each action under the existing network structure.

Table 2. Comparison of indexes and costs under three line-building actions.

Constructed Line	Q Value	C_{in}/M$	C_{EENS}/M$	ζ_{ANRIV}	$f(s_\tau)$/M$
7–8	148.48	3.55	5.51	0.39	39.86
11–14	139.27	4.65	4.63	0.45	42.16
17–18	73.00	3.51	7.72	0.49	46.25

The results in Table 2 show that although the DQN algorithm considers the influence of subsequent actions when calculating Q values, the Q values of line-building actions can also reflect the changes in $f(s_\tau)$ after the construction of line to some extent. This also shows that the agent has a relatively clear judgment on the benefits of the construction of various lines, which can support grid planners to make flexible adjustment of the planning scheme.

The action values of some lines after constructing the three lines with the top three Q values in Figure 10 are shown in Figure 11. Figure 11 shows that the Q value of each action will vary to some extent after the construction of different lines, but they are generally similar (the difference in the Q value of the same action is only 26 at most). The Q value of each action after the construction of line 1–2 in Figure 11 also reflects a problem. After line 1–2 is constructed, the agent cannot adjust the Q value of line 1–2's deleting action well. As a result, line 1–2's deleting action will still have a larger Q value after it is built, which is easy to cause the line to be deleted in the next iteration and enter a loop. Since this paper adds the ε-greedy action selection strategy, the algorithm will jump out of the loop after a small number of iterations to find a feasible scheme. In addition, Figure 11 shows that the Q value of line 1–2 after the construction of itself is the smallest compared with the Q value of line 1–2 after the construction of the other two lines, which indicates that the agent can perceive the change of the network structure to some extent.

Suppose that after taking the Q value of each action and realistic factors into account, line 11–14, 19–20 and 6–10 are constructed in order. Calculate the Q value of each action under the new network structure as before. Excluding the lines that have been built, line 3–24 is the most profitable line-building action. Line 2–8 is selected from other actions with larger Q values, and line 19–20 is selected from actions with smaller Q values. The three lines were constructed respectively and compared with the indexes and costs of scheme 1 as shown in Table 3.

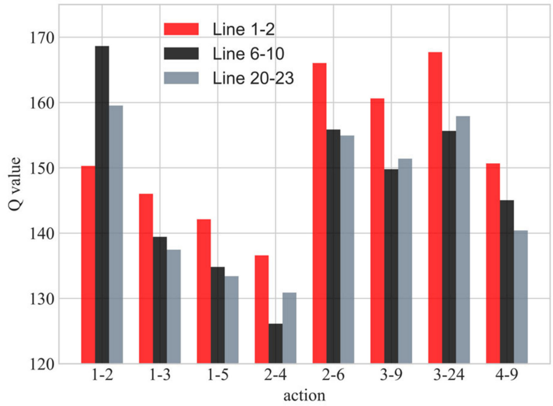


Figure 11. Q value of actions after construction of three different lines.

Table 3. Comparison of indexes and costs of four expansion schemes.

Expansion Scheme	C_{in}/M$	C_{EENS}/M$	ζ_{ANRIV}	$f(s_\tau)$/M$
6–7,19–21,19–22,11–14, 19–20,6–10,3–24	8.94	2.85	0.24	39.20
6–7,19–21,19–22,11–14, 19–20,6–10,2–8	8.22	2.60	0.25	38.22
6–7,19–21,19–22,11–14, 19–20,6–10,19–20	8.70	5.38	0.29	43.72
Scheme 1	6.73	4.27	0.21	37.22

The results in Table 3 show that the selection of line to be constructed according to the magnitude of the action's Q value can make the overall effect of the scheme better to some extent, but it cannot guarantee that the scheme is optimal if the construction is based on the Q value in every step. Therefore, ε-greedy action selection strategy is introduced to make the agent choose other actions so that it can explore better schemes.

5.3. Experiment 3

After the completion of scheme 1, due to the increase in the users' power quality requirements and the increase in load, the current network still cannot meet the users' needs. At this time, the transmission network still needs to be expanded. The proposed method is still able to deal with such problem. For example, line expansion planning should be carried out on the basis of scheme 1. The network structure of scheme 1 should be input into the trained neural network, and the obtained Q value of each action is shown in Figure 12. According to [38], the expansion scheme with the maximum benefit is to build line 10–11. Similarly, line 9–11 is selected from actions with larger Q values, and line 13–23 is selected from actions with smaller Q values. After the construction of three lines, indexes and costs are shown in Table 4.

Table 4. Comparison of indexes and costs under three line-building actions.

Constructed Line	Q Value	C_{in}/M$	C_{EENS}/M$	ζ_{ANRIV}	$f(s_\tau)$/M$
10–11	28.57	8.73	4.90	0.19	39.71
9–11	26.64	8.73	5.67	0.18	40.25
13–23	24.31	10.33	5.57	0.21	41.95

Table 4 shows that although the increase of construction investment cost leads to the increase of $f(s_\tau)$, the two kinds of line-building actions with large Q values can improve the system flexibility. Since the $f(s_\tau)$ of scheme 1 is already relatively small, the reward and Q value of each line-building action on this basis are not large. The Q value of each action in Figure 12 also verifies this conclusion. Therefore, in order to meet the higher demands of the future power system, line 10–11 can be selected for construction.

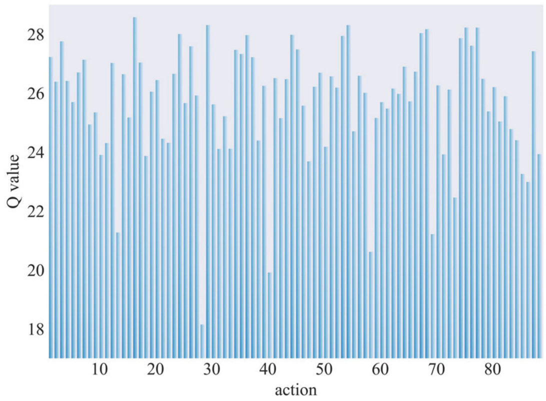


Figure 12. *Q* value of each action after the completion of scheme 1.

6. Conclusions

In this paper, we proposed a TNEP model that comprehensively considers economy, reliability and flexibility. Meanwhile, the possible *N*-k faults and the severity of the equipment faults were taken into account in the Monte Carlo method to calculate the reliability index expected energy not supplied, which was more in line with the actual operating situation. For the implementation process of the expansion scheme, we considered the construction sequence of lines. Compared with mathematical optimization algorithms and meta-heuristic optimization algorithms, the proposed planning method was based on the DQN algorithm, it could solve the multi-stage TNEP problem on the basis of a static TNEP model, and it was able to converge in large-scale systems. In addition, through the repeated use of the trained neural network in the DQN algorithm, the adaptive planning of lines was realized. Moreover, prioritized experience replay strategy was introduced to accelerate the learning efficiency of the agent. Compared with using random sampling strategy, the total number of iterations in the first 50 episodes had been reduced by 30%. Three experiments of the IEEE 24-bus reliability test system showed that the proposed flexible TNEP method could not only complete the multi-stage planning well, but also realize the flexible adjustment of expansion schemes. Selecting the line-building actions based on the *Q* values calculated by the neural network could ensure the justifiability and economy of the obtained scheme to a certain extent.

This study is a first attempt to apply the DQN algorithm to solve TNEP problem considering the construction sequence of lines. Three experiments of the IEEE 24-bus reliability test system verify the effectiveness of the proposed method and its flexibility compared with traditional planning methods. However, this paper is still relatively simple to deal with the multi-stage TNEP problem, and does not consider the specific construction time of each line. In addition, there is an error in the calculation of line-deleting action's *Q* value. How to evaluate the pros and cons of the expansion scheme more comprehensively, consider the multi-stage planning problem more deeply, make the agent have a clearer judgment on the value of line-deleting action, and consider the renewable energy, energy storage, and other equipment are further research directions.

Author Contributions: Conceptualization, Y.W., L.C. and Q.Z.; Data curation, L.J.; Formal analysis, Y.W., Z.Z. and Q.Z.; Funding acquisition, H.Z.; Investigation, Q.Z.; Methodology, L.C. and X.Z.; Project administration, H.Z., L.J. and L.L.; Resources, Y.W. and H.Z.; Software, L.L.; Validation, L.C., X.Z., Z.Z. and L.L.; Visualization, L.J.; Writing—original draft, L.C.; Writing—review & editing, Y.W., X.Z. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Program of State Grid Southwest China Branch (SGSW0000GHJS1900117).

Acknowledgments: The authors would like to thank the editor and reviewers for their sincere suggestions on improving the quality of this paper.

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

Nomenclature

Sets and Indices	
π	Set of strategies
S	Set of states, for example $S = \{l_1, l_2, \ldots, l_n\}$
A	Set of actions
w	Set of feedbacks
W	Set of rewards
Q	Set of Q values
π^*	Set of optimal strategies
Ψ	Set of lines that have been built
$c(i)$	Set of all end buses with i as the head bus
Φ	Set of transmission network operating states
τ	Action index
h	The h-th line
l	Line index
i,j	Bus index
e	Equipment index
Variables	
a_τ	Reinforcement learning agent actions
s_τ	Reinforcement learning states
R_τ	Current action reward
μ	Random number in the interval [0, 1]
d	Current action number
E_π	Mathematical expectation under strategyπ
$Q_\pi(s,a)$	State–action value function
$Q^*(s,a)$	Optimal state–action value function
$\pi(s)$	ε-greedy action selection strategy
Q_{max}	Optimal action's Q value
Q_{eval}	Q value of eval-net
Q_{target}	Q value of target-net
ϖ	eval-net's weight
$L(\varpi)$	eval-net's loss
P_τ	Priority of experience τ
Δ_τ	TD error
$\text{rank}(\lvert\Delta_\tau\rvert)$	Elements of $\lvert\Delta_\tau\rvert$ sorted from the maximum to minimum
$f(s_\tau)$	Comprehensive equivalent cost
C_{in}	Construction investment cost
C_o	Operation and maintenance cost
C_{loss}	Network loss cost
C_{EENS}	Reliability cost
ζ_{ANRIV}	Flexibility index ANRIV
λ_{in}	Annual coefficient of fixed investment cost of the line
β_l	Binary variable, β_l = 1 means line l has been constructed, 0 otherwise
$P_{G,i}$	Total active power output of all generators at bus i
H	Total number of lines between bus i and j
$I_{ij,h}$	Current of the h-th line between bus i and j
$r_{ij,h}$	Resistance of the h-th line between bus i and j
$P_{\text{load},i}$	Active power consumed at bus i
$b_{\text{ij,h}}$	Susceptance of the h-th line between bus i and j

θ_{i}	Voltage phase angle at bus *i*
$P_{\mathrm{G},i}^{\min}$, $P_{\mathrm{G},i}^{\max}$	Lower and upper bound of active power output of all generators at bus *i*
$P_{ij,h}$	Active power flow of the *h*-th line between bus *i* and *j*
$P_{ij,h}^{\max}$	Upper bound of active power flow allowable transmission capacity
U_i	Voltage amplitude at bus *i*
$U_i^{\min}$, $U_i^{\max}$	Lower and upper bound of voltage amplitude at bus *i*
f_z	Load shedding under fault state *z*
$P_{i,z}$	Load shedding at bus *i* under fault state *z*
$P_{G,i,z}$	Sum of the active power output of all generators at bus *i* under fault state *z*
$\theta\ _{j,z}$	Voltage phase angle at bus *i* under fault state *z*
$P_{ij,h,z}$	Active power flow of the *h-th* line between bus *i* and *j* under fault state *z*
δ_e	Operation state of equipment *e*, δ_e=0 represents equipment *e* is out of service
$p_{f,e}$	Forced stop rate of equipment *e*
$P_{\mathrm{G},e}$	Active power output of generator set *e* under normal operation
M_z	Operation state *z* of the transmission network
$p(z)$	Occurrence probability of state *z*
$n(M_z)$	Number of samplings of state *z*
P_{EENS}	Annual load shedding
$P_{i,z}$	Power shortage of bus *i* under state *z*
$\zeta\ _{V,i}$	NRI of the voltage at bus *i*
F	Sum of the number of generators and lines
Constants	
γ	Decay factor of future rewards
D	Upper bound of the action number
α	Learning rate
Γ	Size of the playback experience pool
φ	Degree of priority usage
c_l	Investment cost of unit length construction
ξ	expected return on investment
y_0	Service life of the investment
y_1	Construction life of the planning scheme
n	Number of constructible lines
L_l	Length of line *l*
λ_o	Annual coefficient of operation and maintenance cost of the line
K_G	Unit power generation cost
T	Annual maximum load utilization hours
N	Number of buses
n_{total}	Total number of samplings
K_{loss}	Unit network power loss price
K_{EENS}	Unit load shedding price
η_{ANRIV}	Coefficient that balances the influence of the ζ_{ANRIV} on the objective function

Appendix A

The parameters of the 50 to-be-selected lines are given in Table A1, where the investment costs are converted to the equivalent annual costs and the lines are numbered from 39 to 88. They are used in Section 5, and provide the settings of the experiments. Based on these parameters, the experimental results are analyzed and the planning methods are evaluated.

Table A1. The parameters of the 50 to-be-selected lines.

Number	Line	Cost/M$	Number	Line	Cost/M$
39	1–4	1.12	64	7–9	2.26
40	1–7	2.03	65	7–10	1.72
41	1–8	2.36	66	8–10	1.31
42	1–9	1.75	67	11–15	1.51
43	2–3	2.82	68	11–23	2.40
44	2–5	0.83	69	12–14	1.48
45	2–7	0.86	70	12–15	2.07
46	2–8	1.28	71	13–14	1.76
47	2–9	1.84	72	13–15	2.42
48	2–10	1.72	73	13–20	1.49
49	3–4	1.08	74	14–15	0.67
50	3–5	2.10	75	14–19	0.02
51	3–6	3.22	76	14–20	0.82
52	3–8	3.41	77	15–19	0.93
53	3–10	2.38	78	15–20	1.36
54	4–5	1.08	79	16–18	0.89
55	4–6	2.45	80	16–19	0.89
56	4–10	1.72	81	16–20	1.36
57	5–6	1.56	82	17–19	1.29
58	5–7	1.37	83	18–19	1.11
59	5–8	1.42	84	19–21	0.67
60	5–9	1.02	85	19–22	0.88
61	6–7	1.36	86	20–21	0.80
62	6–8	0.75	87	20–22	0.67
63	6–9	1.82	88	22–23	0.69

References

1. Quintero, J.; Zhang, H.; Chakhchoukh, Y.; Vittal, V.; Heydt, G.T. Next generation transmission expansion planning framework: Models, tools, and educational opportunities. *IEEE Trans. Power Syst.* **2014**, *29*, 1911–1918. [CrossRef]
2. Zhang, X.; Conejo, A.J. Candidate line selection for transmission expansion planning considering long- and short-term uncertainty. *Int. J. Electr. Power Energy. Syst.* **2018**, *100*, 320–330. [CrossRef]
3. Nnachi, G.; Richards, C. A com-prehensive state-of-the-art survey on the transmission network expansion planning optimization algorithms. *IEEE Access* **2019**, *7*, 123158–123181.
4. Garver, L.L. Transmission network estimation using linear programming. *IEEE Trans. Power App. Syst.* **1970**, *PAS-89*, 1688–1697. [CrossRef]
5. Seifu, A.; Salon, S.; List, G. Optimization of transmission line planning including security constraints. *IEEE Trans. Power Syst.* **1989**, *4*, 1507–1513. [CrossRef]
6. Chen, B.; Wang, L. Robust Transmission planning under uncertain generation investment and retirement. *IEEE Trans. Power Syst.* **2016**, *31*, 5144–5152. [CrossRef]
7. Liang, Z.; Chen, H.; Wang, X.; Ibn Idris, I.; Tan, B.; Zhang, C. An extreme scenario method for robust transmission expansion planning with wind power uncertainty. *Energies* **2018**, *11*, 2116. [CrossRef]
8. Zhao, J.H.; Dong, Z.Y.; Lindsay, P.; Wong, K.P. Flexible transmission expansion planning with uncertainties in an electricity market. *IEEE Trans. Power Syst.* **2009**, *24*, 479–488. [CrossRef]
9. Akbari, T.; Rahimikian, A.; Kazemi, A. A multi-stage stochastic transmission expansion planning method. *Energy Convers. Manage.* **2011**, *52*, 2844–2853. [CrossRef]
10. Alizadeh, B.; Jadid, S. Reliability constrained coordination of generation and transmission expansion planning in power systems using mixed integer programming. *IET Gener. Transm. Distrib.* **2011**, *5*, 948–960. [CrossRef]
11. Feng, Z.K.; Niu, W.J.; Cheng, C.T.; Liao, S.L. Hydropower system operation optimization by discrete differential dynamic programming based on orthogonal experiment design. *Energy* **2017**, *126*, 720–732. [CrossRef]
12. Zakeri, A.; Abyaneh, H. Transmission expansion planning using TLBO algorithm in the presence of demand response resources. *Energies* **2017**, *10*, 1376. [CrossRef]
13. Gallego, L.A.; Garcés, L.P.; Rahmani, M.; Romero, R.A. High-performance hybrid genetic algorithm to solve transmission network expansion planning. *IET Gener. Transm. Distrib.* **2017**, *11*, 1111–1118. [CrossRef]
14. Fuerte Ledezma, L.F.; Gutiérrez Alcaraz, G. Hybrid binary PSO for transmission expansion planning considering N-1 security criterion. *IEEE Lat. Am. Trans.* **2020**, *18*, 545–553. [CrossRef]

15. Cheng, L.; Yu, T. A new generation of AI: A review and perspective on machine learning technologies applied to smart energy and electric power systems. *Int. J. Energ. Res.* **2019**, *43*, 1928–1973. [CrossRef]
16. Wang, J.; Tao, Q. Machine learning: The state of the art. *IEEE Intell. Syst.* **2008**, *23*, 49–55. [CrossRef]
17. Hu, W.; Zheng, L.; Min, Y.; Dong, Y.; Yu, R.; Wang, L. Research on power system transient stability assessment based on deep learning of big data technique. *Power Syst. Technol.* **2017**, *41*, 3140–3146.
18. Cheng, L.; Yu, T. Dissolved gas analysis principle-based intelligent approaches to fault diagnosis and decision making for large oil-immersed power transformers: A survey. *Energies* **2018**, *11*, 913. [CrossRef]
19. Ryu, S.; Noh, J.; Kim, H. Deep neural network based demand side short term load forecasting. *Energies* **2017**, *10*, 3. [CrossRef]
20. Tan, M.; Yuan, S.; Li, S.; Su, Y.; Li, H.; He, F. Ultra-short-term industrial power demand forecasting using LSTM based hybrid ensemble learning. *IEEE Trans. Power Syst.* **2020**, *35*, 2937–2948. [CrossRef]
21. Xiang, L.; Zhao, G.; Li, Q.; Hao, W.; Li, F. TUMK-ELM: A fast unsupervised heterogeneous data learning approach. *IEEE Access* **2018**, *7*, 35305–35315. [CrossRef]
22. Littman, M.L. Reinforcement learning improves behaviour from evaluative feedback. *Nature* **2015**, *521*, 445–451. [CrossRef]
23. Silver, D.; Huang, A.; Maddison, C.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schritwieser, J.; Antonoglou, L.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef]
24. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellermare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
25. Ji, Y.; Wang, J.; Xu, J.; Fang, X.; Zhang, H. Real-time energy management of a microgrid using deep reinforcement learning. *Energies* **2019**, *12*, 2291. [CrossRef]
26. Yu, T.; Wang, H.Z.; Zhou, B.; Chan, K.W.; Tang, J. Multi-agent correlated equilibrium Q(λ) learning for coordinated smart generation control of interconnected power grids. *IEEE Trans. Power Syst.* **2015**, *30*, 1669–1679. [CrossRef]
27. Xi, L.; Yu, L.; Xu, Y.; Wang, S.; Chen, X. A novel multi-agent DDQN-AD method-based distributed strategy for automatic generation control of integrated energy systems. *IEEE Trans. Sustain. Energy* **2020**, *11*, 2417–2426. [CrossRef]
28. Cao, J.; Zhang, W.; Xiao, Z.; Hua, H. Reactive power optimization for transient voltage stability in energy internet via deep reinforcement learning approach. *Energies* **2019**, *12*, 1556. [CrossRef]
29. Hong, S.; Cheng, H.; Zeng, P. An N-k analytic method of composite generation and transmission with interval load. *Energies* **2017**, *10*, 168. [CrossRef]
30. Zhang, Y.; Wang, J.; Li, Y.; Wang, X. An extension of reduced disjunctive model for multi-stage security-constrained transmission expansion planning. *IEEE Trans. Power Syst.* **2018**, *33*, 1092–1094. [CrossRef]
31. Kim, W.-W.; Park, J.-K.; Yoon, Y.-T.; Kim, M.-K. Transmission expansion planning under uncertainty for investment options with various lead-times. *Energies* **2018**, *11*, 2429. [CrossRef]
32. Arabali, A.; Ghofrani, M.; Etezadi-Amoli, M.; Fadali, M.S.; Moeini-Aghtaie, M. A multi-objective transmission expansion planning framework in deregulated power systems with wind generation. *IEEE Trans. Power Syst.* **2014**, *29*, 3003–3011. [CrossRef]
33. Kamyab, G.-R.; Fotuhi-Firuzabad, M.; Rashidinejad, M. A PSO based approach for multi-stage transmission expansion planning in electricity markets. *Int. J. Electr. Power Energy Syst.* **2014**, *54*, 91–100. [CrossRef]
34. Qiu, J.; Zhao, J.; Wang, D. Flexible multi-objective transmission expansion planning with adjustable risk aversion. *Energies* **2017**, *10*, 1036. [CrossRef]
35. Schaul, T.; Quan, J.; Antonoglou, I.; Silver, D. Prioritized experience replay. In Proceedings of the International Conference on Learning Representations 2016, San Juan, Puerto Rico, 2–4 May 2016.
36. Ni, M.; McCalley, J.D.; Vittal, V.; Tayyib, T. Online risk-based security assessment. *IEEE Trans. Power Syst.* **2003**, *18*, 258–265. [CrossRef]
37. Subcommittee, P.M. IEEE Reliability Test System. *IEEE Trans. Power App. Syst.* **1979**, *PAS-98*, 2047–2054. [CrossRef]
38. Zhang, H.; Vittal, V.; Heydt, G.T.; Quintero, J. A mixed-integer linear programming approach for multi-stage security-constrained transmission expansion planning. *IEEE Trans. Power Syst.* **2012**, *27*, 1125–1133. [CrossRef]