electronics

Article

Managing Considerable Distributed Resources for Demand Response: A Resource Selection Strategy Based on Contextual Bandit

Zhaoyu Li and Qian Ai *

Key Laboratory of Control of Power Transmission and Conversion, Ministry of Education, Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; zhaoyuli72@sjtu.edu.cn
* Correspondence: aiqian@sjtu.edu.cn

Abstract: The widespread adoption of distributed energy resources (DERs) leads to resource redundancy in grid operation and increases computation complexity, which underscores the need for effective resource management strategies. In this paper, we present a novel resource management approach that decouples the resource selection and power dispatch tasks. The resource selection task determines the subset of resources designated to participate in the demand response service, while the power dispatch task determines the power output of the selected candidates. A solution strategy based on contextual bandit with DQN structure is then proposed. Concretely, an agent determines the resource selection action, while the power dispatch task is solved in the environment. The negative value of the operational cost is used as feedback to the agent, which links the two tasks in a closed-loop manner. Moreover, to cope with the uncertainty in the power dispatch problem, distributionally robust optimization (DRO) is applied for the reserve settlement to satisfy the reliability requirement against this uncertainty. Numerical studies demonstrate that the DQN-based contextual bandit approach can achieve a profit enhancement ranging from 0.35% to 46.46% compared to the contextual bandit with policy gradient approach under different resource selection quantities.

Keywords: contextual bandit; resource selection; demand response; distributionally robust optimization

check for updates

Citation: Li, Z.; Ai, Q. Managing Considerable Distributed Resources for Demand Response: A Resource Selection Strategy Based on Contextual Bandit. *Electronics* **2023**, *12*, 2783. https://doi.org/10.3390/electronics12132783

Academic Editor: Adel M. Sharaf

Received: 9 May 2023
Revised: 21 June 2023
Accepted: 21 June 2023
Published: 23 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background and Motivation

The proliferation of distributed energy resources (DERs) is anticipated within regional energy networks. For instance, the total installed capacity of wind power and solar power generation in China will exceed 1.2 billion kW by 2030 [1]. Due to the ongoing concern about climate change, China is committed to achieving a CO_2 emissions peak before 2030, and carbon neutrality by 2060 [2], which underscores a pressing need for DERs to replace fossil fuels and participate in regional energy network operations. The integration of DERs in this manner can offer a range of benefits, such as meeting regional electricity demands through local resources [3] and providing energy support for the upper grid [4].

However, simply allowing a significant number of resources to participate in the operation without appropriate selection may result in challenges. For instance, this could lead to redundancy, whereby a considerable number of resources may have zero power exchange during the scheduling horizon, which could dampen their initiatives. Thus, it is crucial to select the optimal portfolio of resources from the entire set of resources, which falls under the realm of combinatorial optimization, and can be formulated as a mixed-integer linear programming (MILP) problem [5]. Although such a MILP problem can be solved by the branch-and-bound technique using existing commercial solvers, this approach can be computationally challenging when the number of binary variables increases [5]. Moreover, the practice in [6,7] requires the binary and continuous variables to be solved by a single entity. However, for resource management problems, the resources selection task and the

power dispatch task may need to be solved by distinct entities. Therefore, an efficient resource selection approach that is applicable when confronted with a large number of resources and that can accommodate distinct entities is needed.

Therefore, in this paper, we consider a resource management problem in which considerable flexible resources can participate in load balance and demand response service. Instead of scheduling all of the resources, only a subset of the resources are chosen. The challenge of this problem lies in the coupling between the upstream resource selection problem and the downstream energy dispatch problem.

1.2. Literature Review and Research Gap

Consisting of the intercorrelated elements of agent and environment, reinforcement learning (RL) is a promising approach for solving problems with complex coupling between two decision-making processes [8,9]. RL has been successfully applied in power scheduling [10], demand response strategy learning [11], energy pricing and bidding [12,13], residential resource control [14], etc. However, in a resource management problem, the state is not affected by selection, and hence, the application of RL techniques that involve state evolution is not straightforward. As an alternative, the contextual bandit approach offers a more suitable framework by mapping the features (i.e., the context) to the action. Contextual bandit can be considered a one-step RL, wherein the optimal strategy is determined solely based on the current context, rather than considering the overall strategy that considers the dynamic evolution of the system [15]. Algorithms developed for solving RL problems, such as policy gradient methods, can be adapted for contextual bandit problems [16].

Currently, many efforts have been devoted to dealing with sequential decision-making problems using bandit-based approaches, where the decision maker seeks to select the action with the highest expected reward among action candidates under uncertainty [17]. Applications have been applied to several fields, for instance, recommendation systems [18], information retrieval [19], healthcare [20], etc. Several current pieces of research about the application of bandit-based approaches to the energy resource management problem are provided in Table 1. Multi-armed bandit (MAB) approaches were applied to learn the behaviors of residential air conditioning [21]; heating, ventilation, and air conditioning (HVAC) [22]; renewable energy sources [23]; and energy storage [24], selecting the optimal set to participate in primary or secondary frequency regulation. Moreover, the optimal set of electric vehicles [25] and residential demand resources [26,27] were selected based on MAB-based approaches to provide demand response services to the power grid operation. However, the aforementioned approaches mainly focus on the uncertainty pertaining to the participation of a singular resource type, and the selected resources are aggregated to ensure the cumulative capacity adequately satisfies the predetermined demand. However, limited consideration has been paid to the operational impact of resource selection decisions for multiple types of resources, particularly in the context of uncertain demand requirements. Therefore, this paper employs the contextual bandit approach to effectively enable the selection of multiple resources in the presence of operational uncertainties.

Table 1. Applications of bandit-based approaches.

Ref.	Resource Type	Service	Approach
[21]	Residential air conditioning	Primary and secondary frequency regulation	Risk-averse MAB
[22]	Heating, ventilation, and air conditioning (HVAC)	Secondary frequency regulation	Risk-averse MAB
[23]	Renewable Energy Sources	Secondary frequency regulation	MAB
[24]	Energy storage	Primary frequency regulation	MAB
[25]	Electric vehicle	Ancillary services	MAB
[26]	Residential demand	Demand response	MAB
[27]	Residential demand	Demand response	Contextual MAB
This paper	Diesel generator, Gas turbine, Curtailable load	Load balance, Demand response	Contextual MAB

Stochastic programming and robust optimization are two approaches typically used to address uncertainty in an operation. Stochastic programming can be further divided into chance-constrained programming (CCP) [28] and scenario-based approaches [29]. Although stochastic programming has wide application in unit commitment [30], long-term planning [31], etc., it has some limitations. For instance, the assumed probability distribution of CCP may not be accurate in some cases, and the number of generated scenarios required by a scenario-based approach may cause a computational burden. Moreover, robust optimization makes decisions based on the worst-case scenario of uncertainty, leading to overly conservative solutions [32]. Distributionally robust optimization (DRO), which combines CCP and robust optimization, shows promise in overcoming the drawbacks of these two approaches. With no assumption on the probability distribution of uncertainty, DRO can obtain a less conservative solution than robust optimization [33].

The formulation of the uncertainty set for DRO is very critical. At present, there are two common types of methods to construct uncertainty sets. The first type is the metric-based approach [34], which constructs uncertainty sets for distributions based on statistical distances. Commonly used statistical distances include Prohorov [35], Kullback–Leibler scatter [36], Wasserstein distance [37], and ϕ-scatter [38], etc. The other type is the moment-based approach, which utilizes moment information (specifically, the first-order and the second-order moments) to construct uncertainty sets [39]. Since the moment-based approach can be conveniently reformulated into tractable deterministic optimization problems, it is suitable for addressing uncertainty in a power dispatch problem.

1.3. Contribution and Organization

In this paper, we decouple the resource management problem in a demand response service into two tasks, namely, resource selection and power dispatch. Among a large number of resources, the resource selection task determines which resources are eligible to participate in the demand response service, while the energy dispatch task solves the energy dispatch (ED) problem to determine the optimal power output for the selected resources. Since the two tasks interact with each other, a contextual bandit approach with a deep Q Network (DQN) structure is leveraged to learn their complex interdependency. Concretely, given the resources' features as the input context, the agent learns the optimal policy of resource selection based on a DQN. Then, given the action of participant selection made by the agent, the power dispatch task determines the optimal power dispatch among the selected participants in the environment under the contextual bandit's framework. The operational cost of the power dispatch task is treated as the reward and passed back to guide the agent's selection decision, thereby forming a closed-loop manner. Additionally, the uncertainty in the decision-making process of power dispatch is handled with DRO, and distributionally robust variants for the reserve capacity constraints are formulated. The main contributions of the paper are summarized as follows:

(1) A new resource management model, which decouples the resource selection problem and the power dispatch problem to avoid redundancy and unnecessary participation.
(2) A solution strategy based on contextual bandit with a DQN structure, which treats the performance of the power dispatch problem as a reward, and learns the policy of resource selection in a closed-loop manner.
(3) A distributionally robust variant to cope with the demand uncertainty in the power dispatch task, whose superiorities are empirically demonstrated.

The remainder of this paper is organized as follows. Section 2 introduces the proposed framework and the power dispatch problem. Section 3 illustrates the proposed solution strategy based on the contextual bandit framework. Results are discussed and evaluated in Section 4. Section 5 provides our conclusion.

2. Problem Formulation

2.1. The Architecture of the Regional Energy System

The proposed architecture for the resource management problem is shown in Figure 1. The regional energy system comprises a large number of distributed resources that can be utilized to balance the local demand and participate in the demand response program. The distribution company (DisCo) is responsible for selecting the appropriate resources based on the predicted load level and demand response requirement. Instead of allowing all available resources to participate, the resource management problem comprises two interrelated tasks, namely, resource selection and power dispatch. Concretely, in the resource selection task, DisCo selects the best portfolio, which is eligible to fulfill the service, based on contextual information such as the predicted load and demand response requirement. Subsequently, based on the selection decision, the optimization problem of power dispatch, which only involves the continuous variables, is performed among the selected resources in the regional energy network. Therefore, the resource management problem involves the complex coupling between the two tasks.

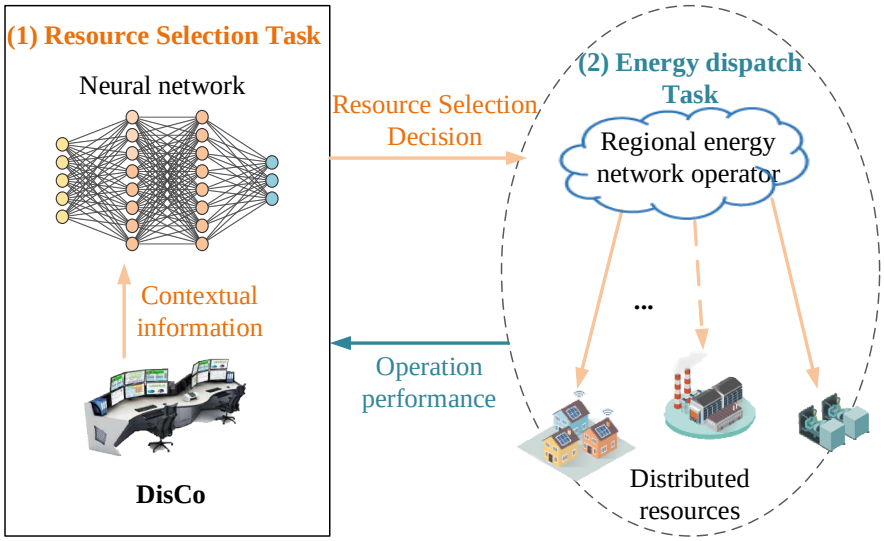

Figure 1. The illustration of the proposed architecture for the regional energy system.

The compact mathematical form of the resource management problem at a single time slot can be formulated as follows.

$$\min_{x,y} \boldsymbol{c}^T \boldsymbol{y} \tag{1a}$$

$$s.t.\ \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} \leq \boldsymbol{g} \tag{1b}$$

$$\boldsymbol{H}\boldsymbol{x} \cdot \boldsymbol{M}\boldsymbol{y} = \boldsymbol{d} \tag{1c}$$

$$\mathbf{1}^T \boldsymbol{x} = K \tag{1d}$$

$$\boldsymbol{x}\ binary \tag{1e}$$

where $\boldsymbol{c}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{H}, \boldsymbol{M}$ and $\boldsymbol{g}, \boldsymbol{d}$ are the coefficient matrixes and the coefficient vectors, separately. The all one vector is denoted by $\mathbf{1}$. The binary variable $\boldsymbol{x}$ represents the selection decision, which indicates the operation statuses of the resources. The continuous variable $\boldsymbol{y}$ refers to the power dispatch schedule of the resources. The overall objective is to minimize the management cost while satisfying the predicted load and demand response requirement. Constraints (1b) and (1c) couple the decisions of the resource selection problem and the power dispatch problem. Constraint (1b) gives the generation capacity and load curtailment limitation constraints, and constraint (1c) represents the energy balance constraint. Constraint (1d) restricts the number of the selected resources to a given integer K.

2.2. The Power Dispatch Problem

This section establishes the mathematical formulation of the power dispatch problem. The resources considered include diesel generators (DGs), gas turbines (GTs), and curtailable loads.

2.2.1. Diesel Generator

The output power of DGs consists of two parts: power balancing of the predicted load and participation in the demand response service, respectively. Let $\mathcal{A}$ denote the set of DGs, such that $i \in \mathcal{A}$ represents the index of the DGs analyzed. The constraints of DGs can be formulated as follows:

$$0 \le P_{i,t}^{dg} \le \overline{P}_i^{dg} \cdot B_i^{dg} \tag{2}$$

$$0 \le P_{i,t}^{dg} + P_{i,t}^{dg,DR} \le \overline{P}_i^{dg} \cdot B_i^{dg} \tag{3}$$

$$-R_i^{dg,D} \cdot B_i^{dg} \le P_{i,t+1}^{dg} - P_{i,t}^{dg} \le R_i^{dg,U} \cdot B_i^{dg} \tag{4}$$

$$-R_i^{dg,D} \cdot B_i^{dg} \le P_{i,t+1}^{dg} + P_{i,t+1}^{dg,DR} - P_{i,t}^{dg} - P_{i,t}^{dg,DR} \le R_i^{dg,U} \cdot B_i^{dg} \tag{5}$$

$$-\overline{P}_i^{dg} \cdot B_i^{dg} \le P_{i,t}^{dg,DR} \le \overline{P}_i^{dg} \cdot B_i^{dg} \tag{6}$$

where B_i^{dg} is the binary variable determined by the DisCo in the resource selection problem, which is a parameter in the power dispatch problem. When $B_i^{dg} = 1$, the i-th DG is chosen to balance the predicted load and participate in the demand response service. When $B_i^{dg} = 0$, the i-th DG is not chosen. Constraints (2) and (3) restrict the output power of DG within the allowable range. Constraints (4) and (5) give the lower and upper bounds of the ramp constraints. Constraint (6) limits the output power of DG in the demand response service.

2.2.2. Gas Turbine

The outputs of GTs are leveraged to balance the local demand and serve in the demand response program. We assume that $\mathcal{B}$ denotes the set of GTs, and $i \in \mathcal{B}$ represents the index of chosen GTs. The constraints of the GTs can be formulated as follows:

$$0 \le P_{i,t}^{gt} \le \overline{P}_i^{gt} \cdot B_i^{gt} \tag{7}$$

$$0 \le P_{i,t}^{gt} + P_{i,t}^{gt,DR} \le \overline{P}_i^{gt} \cdot B_i^{gt} \tag{8}$$

$$-R_i^{gt,D} \cdot B_i^{gt} \le P_{i,t+1}^{gt} - P_{i,t}^{gt} \le R_i^{gt,U} \cdot B_i^{gt} \tag{9}$$

$$-R_i^{gt,D} \cdot B_i^{gt} \le P_{i,t+1}^{gt} + P_{i,t+1}^{gt,DR} - P_{i,t}^{gt} - P_{i,t}^{gt,DR} \le R_i^{gt,U} \cdot B_i^{gt} \tag{10}$$

$$-\overline{P}_i^{gt} \cdot B_i^{gt} \le P_{i,t}^{gt,DR} \le \overline{P}_i^{gt} \cdot B_i^{gt} \tag{11}$$

where B_i^{gt} is the determined binary variable in the resource selection problem, which is treated as a parameter in the power dispatch problem. When $B_i^{gt} = 1$, the i-th GT is chosen to generate power. When $B_i^{gt} = 0$, the i-th GT is not chosen. Constraints (7) and (8) limit the lower and upper bounds of the GTs' output. Constraints (9) and (10) represent the ramp limitation of the GTs. Constraint (11) shows the range of the output power of the GTs in the demand response service.

2.2.3. Curtailable Load

Let $\mathcal{C}$ denote the set of curtailable loads, such that $i \in \mathcal{C}$ represents the index of the curtailable loads analyzed.

$$0 \le P_{i,t}^{cu} \le \overline{P}_i^{cu} \cdot B_i^c \tag{12}$$

Constraint (12) restricts the range of the curtailable load. B_i^c is a DisCo determined parameter in the power dispatch problem. When $B_i^c = 1$, the i-th curtailable load is chosen in the demand response service. When $B_i^c = 0$, it is not chosen.

Therefore, the mathematical formulation of the power dispatch problem is:

$$\min c = \sum_{t \in T} P_t^{grid} \cdot \pi_t + \sum_{i \in \mathcal{A}} c_{dg} + \sum_{i \in \mathcal{B}} c_{gt} + \sum_{i \in \mathcal{C}} c_c \tag{13}$$

$$c_{dg} = \sum_{t \in T} \left[\rho_i^{dg} \left(P_{i,t}^{dg} + P_{i,t}^{dg,DR} \right) - \pi^{DR} \cdot P_{i,t}^{dg,DR} \right] \cdot B_i^{dg} \tag{14}$$

$$c_{gt} = \sum_{t \in T} \left[\pi^{gas} \cdot \eta_i^{gt} \left(P_{i,t}^{gt} + P_{i,t}^{gt,DR} \right) - \pi^{DR} \cdot \eta_i^{gt} \cdot P_{i,t}^{gt,DR} \right] \cdot B_i^{gt} \tag{15}$$

$$c_c = \sum_{t \in T} \left(-\pi^{DR} \cdot P_{i,t}^{cu} \right) \cdot B_i^c \tag{16}$$

It has the following system constraints:

$$P_t^{grid} + \sum_{i \in \mathcal{A}} P_{i,t}^{dg} \cdot B_i^{dg} + \sum_{i \in \mathcal{B}} P_{i,t}^{gt} \cdot B_i^{gt} = P_t^{load,P} + \sum_{i \in \mathcal{C}} \overline{P}_i^{cu} - \sum_{i \in \mathcal{C}} P_{i,t}^{cu} \cdot B_i^c \tag{17}$$

$$\sum_{i \in \mathcal{A}} P_{i,t}^{dg,DR} \cdot B_i^{dg} + \sum_{i \in \mathcal{B}} P_{i,t}^{gt,DR} \cdot B_i^{gt} + \sum_{i \in \mathcal{C}} P_{i,t}^{cu} \cdot B_i^c = P_t^{DR} \tag{18}$$

$$P\left(\sum_{i \in \mathcal{A}} \overline{P}_i^{dg} \cdot B_i^{dg} - \sum_{i \in \mathcal{A}} P_{i,t}^{dg} \cdot B_i^{dg} + \sum_{i \in \mathcal{B}} \overline{P}_i^{gt} \cdot B_i^{gt} - \sum_{i \in \mathcal{B}} P_{i,t}^{gt} \cdot B_i^{gt} \ge \delta_t^{load} \right) \ge 1 - \alpha \tag{19}$$

$$P\left(\sum_{i \in \mathcal{C}} \overline{P}_i^{cu} \cdot B_i^c - \sum_{i \in \mathcal{C}} P_{i,t}^{cu} \cdot B_i^c \ge \delta_t^{DR} \right) \ge 1 - \alpha \tag{20}$$

where the cost in (13) is minimized over a single time period indexed by $t \in T$. The first term of (13) is the cost of purchasing power from the grid. On the right side of function (14), the first term indicates the operating cost of the DGs, while the second term denotes the revenue obtained from participating in the demand response service. Likewise, the GTs' costs for and profits from participating in the demand response service are presented in (15). Equation (16) presents the compensation for the load curtailment. Constraint (17) indicates that the total generation should be equal to the total load demand. Additionally, constraint (18) ensures that the demand response requirement is met through the output of the DGs, GTs, and load curtailment.

Since the predicted load and demand response requirement have uncertainty, some DG, GT, and curtailable load capacities are spared to improve the reliability with a certain probability. The chance constraints are formulated in (19) and (20). $\delta_t^{load}, \delta_t^{DR}$ are the corresponding random variables of the estimation errors and α is the risk parameter. Then, the distributionally robust variants for the chance constraints are:

$$\inf_{f(\delta_t^{load}) \in D_t^{load}} P_{\delta_t^{load}} \left(\sum_{i \in \mathcal{A}} P_{i,t}^{dg} \cdot B_i^{dg} - \sum_{i \in \mathcal{A}} \overline{P}_i^{dg} \cdot B_i^{dg} + \sum_{i \in \mathcal{B}} P_{i,t}^{gt} \cdot B_i^{gt} - \sum_{i \in \mathcal{B}} \overline{P}_i^{gt} \cdot B_i^{gt} + \delta_t^{load} \le 0 \right) \tag{21}$$

$$\inf_{f(\delta_t^{DR}) \in D_t^{DR}} P_{\delta_t^{DR}} \left(\sum_{i \in \mathcal{C}} P_{i,t}^{cu} \cdot B_i^c - \sum_{i \in \mathcal{C}} \overline{P}_i^{cu} \cdot B_i^c + \delta_t^{DR} \le 0 \right) \tag{22}$$

where D_t^{load}, D_t^{DR} and $f\left(\delta_t^{load}\right), f\left(\delta_t^{DR}\right)$ are the ambiguity sets and probability density functions of $\delta_t^{load}, \delta_t^{DR}$, respectively. The ambiguity set is defined by the first and second moments [40]:

$$D = \left\{ f(\xi): \begin{array}{c} \int f(\xi)d\xi = 1 \\ E[\xi] = 0 \\ E\left[\xi^2\right] = \sigma^2 \end{array} \right. \tag{23}$$

Since the random variables $\delta_t^{load}, \delta_t^{DR}$ have the same type of ambiguity set, for simplicity, we use the variable ξ to denote those variables in (23).

Through the deterministic convex reformulation [40], the distributionally robust constraints (21), (22) can be rewritten as:

$$\sum_{i\in\mathcal{A}} P_{i,t}^{dg}\cdot B_i^{dg} - \sum_{i\in\mathcal{A}} \overline{P}_i^{dg}\cdot B_i^{dg} + \sum_{i\in\mathcal{B}} P_{i,t}^{gt}\cdot B_i^{gt} - \sum_{i\in\mathcal{B}} \overline{P}_i^{gt}\cdot B_i^{gt} + \sqrt{\frac{1-\alpha}{\alpha}}\cdot\sigma_t^{load} \le 0 \tag{24}$$

$$\sum_{i\in\mathcal{C}} P_{i,t}^{cu}\cdot B_i^{c} - \sum_{i\in\mathcal{C}} \overline{P}_i^{cu}\cdot B_i^{c} + \sqrt{\frac{1-\alpha}{\alpha}}\cdot\sigma_t^{DR} \le 0 \tag{25}$$

Therefore, we can solve the power dispatch problem as follows:

$$\begin{aligned} &\min c = \sum_{t\in T} P_t^{grid}\cdot\pi_t + \sum_{i\in\mathcal{A}} c_{dg} + \sum_{i\in\mathcal{B}} c_{gt} + \sum_{i\in\mathcal{C}} c_c \\ &s.t. (2)\text{–}(12), (14)\text{–}(18), (24)\text{–}(25) \end{aligned} \tag{26}$$

3. Solution Strategy

3.1. The Contextual Bandit Formulation

We design a closed-loop framework between the resource selection problem and the power dispatch problem using the contextual bandit algorithm. The resource selection problem needs to determine K participants out of the $|\mathcal{A}| + |\mathcal{B}| + |\mathcal{C}|$ distributed resources to balance the predicted load and participate in the demand response service, where K is the parameter set by DisCo. Given the participant selection decision, the power dispatch problem solves the problem in (26). The cost obtained from Equation (13) reflects the quality of the decision made in the resource selection problem, and affects the decision-making process in turn. The illustration of the contextual bandit-based resource selection framework is depicted in Figure 2.

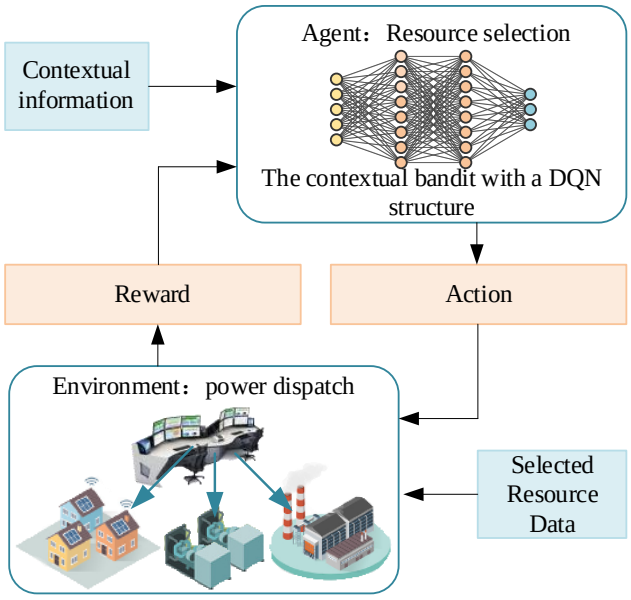

Figure 2. The illustration of the proposed architecture for the regional energy system.

The detailed formulation of the problem and the key elements are outlined in the following.

Agent: the agent learns the policy that determines which resources to select for participation in the demand response program. The policy is learned based on a deep Q-network structure that maps contextual information to specific actions.

Environment: the optimal power dispatch problem of regional distributed resources is solved in the environment.

Context: the context vector plays the role of input feature for the agent regarding the demand level of predicted load and demand response service. Here, this is a $2\cdot|T|$-dimensional vector that can reflect the predicted load and demand response requirement. Thus, it has the form of $\left[P_1^{load,P};\ldots;P_{|T|}^{load,P};P_1^{DR};\ldots;P_{|T|}^{DR}\right]$.

Action: the output action $\boldsymbol{a}$ is a vector that has the form of $\left[B_1^{dg},\ldots,B_{|\mathcal{A}|}^{dg},B_1^{gt},\ldots,B_{|\mathcal{B}|}^{gt},B_1^{c},\ldots,B_{|\mathcal{C}|}^{c}\right]^T$. If the element a_i of the action vector equals one, the corresponding resource is selected as a participant. Conversely, if the element a_i of the action vector equals zero, the corresponding resource is not selected.

Reward: If (26) is feasible and can be solved, the reward is set as the negative value of the operational cost in (13):

$$R(\boldsymbol{a}) = -c/c_{base} \tag{27}$$

where c_{base} is the constant base profit.

If the chosen action causes (26) to be unsolvable, the reward is set to a small negative constant. In our case, we set $R(\boldsymbol{a})$ as -5. Therefore, the agent can learn the policy of resource selection to maximize the reward, which in turn minimizes the cost.

3.2. The Contextual Bandit with a DQN Structure

The flowchart of the contextual bandit approach with a DQN structure is depicted in Figure 3. Based on the training set, the agent learns the state-action approximation function $Q(\boldsymbol{s}_t,\boldsymbol{a}_t|\boldsymbol{\theta})$, which approximates the reward value r_t to minimize the estimation error. The loss function is defined as (28).

$$L = \min_{\boldsymbol{\theta}} \sum_{t\in\mathcal{T}^{tr}} \frac{1}{2}\left(Q(\boldsymbol{s}_t,\boldsymbol{a}_t|\boldsymbol{\theta}) - r_t\right)^2 \tag{28}$$

The DNN parameter $\hat{\boldsymbol{\theta}}$ is updated by gradient descent. By randomly sampling a batch of data $B^Q = \{\boldsymbol{s}_t,\boldsymbol{a}_t\}_{t=1}^{B}$ from the agent's buffer D^Q, the parameter is updated based on (29).

$$\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}} - \eta_Q \cdot \sum_{t\in B^Q} \left(Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t|\hat{\boldsymbol{\theta}}\right) - r_t\right) \frac{\partial Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t|\hat{\boldsymbol{\theta}}\right)}{\partial\hat{\boldsymbol{\theta}}} \tag{29}$$

where, η_Q is the learning rate. Based on the estimated parameter $\hat{\boldsymbol{\theta}}$, the ϵ-greedy algorithm is used to select actions. To make a balance between exploration and exploitation, the agent chooses the best action $\boldsymbol{a}_t = \arg\max Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t|\hat{\boldsymbol{\theta}}\right)$ with a probability of $1-\epsilon$, and randomly chooses an action from the action space with a probability of ϵ. In the learning process, the explore rate ϵ decays exponentially from the maximization value $\epsilon_{\max}$ to the minimization value $\epsilon_{\min}$, which is defined as:

$$\varphi(e,\epsilon) = \left(\epsilon_{\text{decay}}\right)^{e-1}\cdot\epsilon_{\max} \tag{30}$$

where e is the number of epochs. Therefore, based on the state-action approximation function $Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t|\hat{\boldsymbol{\theta}}\right)$ and the ϵ-greedy algorithm, the resource selection decision is determined. The pseudocode of the algorithm is presented in Algorithm 1.

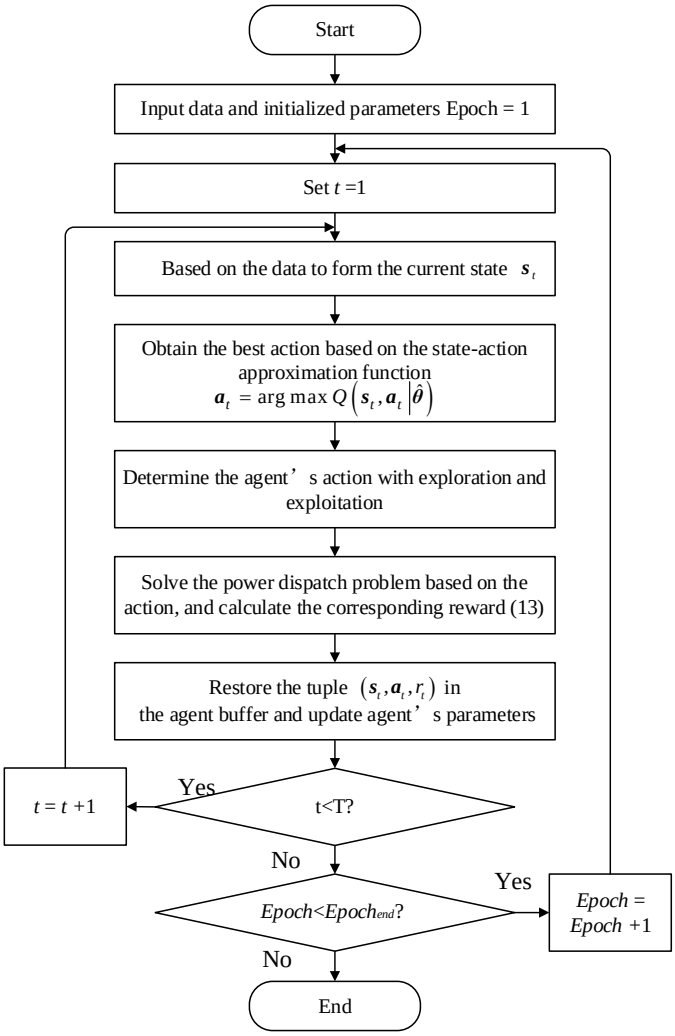

Figure 3. The Flowchart of the contextual bandit approach with a DQN structure.

Algorithm 1: Contextual bandit approach with a DQN structure

1. **Input**: Batch size B, learning rate η_Q, $epoch = E$, learning parameters $\epsilon_{\min}$,$\epsilon_{\max}$,ϵ_{decay}.
2. **Initialize**: The parameter $\boldsymbol{\theta}$ of the DQN-based agent is initialized randomly.
3. **for** $e = 1 : E$ **do**
4. **for** $t = 1 : \left|\mathcal{T}^{tr}\right|$ **do**
5. Based on the input state $\boldsymbol{s}_t$, the agent selects random action $\boldsymbol{a}_t$ with probability ϵ, otherwise selects the best action $\boldsymbol{a}_t = \text{argmax}Q\left(\boldsymbol{s}_t, \boldsymbol{a}_t \middle| \hat{\boldsymbol{\theta}}\right)$.
6. Execute the action $\boldsymbol{a}_t$ in the environment: Solve the power dispatch problem (26), and obtain the reward according to (13).
7. Restore the tuple $(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t)$ in the agent buffer D^Q.
8. Randomly sample dataset B^Q with batch size of B from the buffer D^Q, and update the agent's network parameter according to (29).
9. **if** $\epsilon > \epsilon_{\min}$
10. $\epsilon \leftarrow \varphi(e, \epsilon)$
11. **else**
12. $\epsilon \leftarrow \epsilon_{\min}$
13. **end**
14. **end for**
15. **end for**

4. Case Study

4.1. Implementation Detail

The regional energy network studied is composed of one hundred DGs, one hundred GTs, and fifty curtailable loads, which are available for local demand balance and to

participate in the demand response service. The load data obtained from the Low Carbon London Trail (LCL) [41] is utilized in this case. The cost is calculated in the currency of Chinese RMB.

The model of DGs is formulated as (2)–(6), (14). The marginal costs of DGs are distributed between [0.2,0.4] RMB/kWh. The ramp up parameters $R_i^{dg,U}$, ramp down parameters $R_i^{dg,D}$ and the maximum output powers $\overline{P}_i^{dg}$ of DGs are distributed between [30, 50] kW, [20, 40] kW, and [80, 120] kW, respectively. The model of GTs is developed as (7)–(11), (15), and the generation efficiencies of GTs are distributed between [0.4, 0.6] km^3/kWh. The ramp up parameters $R_i^{gt,U}$, ramp down parameters $R_i^{gt,D}$ and the maximum output powers $\overline{P}_i^{gt}$ of GTs are distributed between [50, 60] kW, [30, 50] kW, and [80, 100] kW, respectively. For curtailable loads, the model is built as (12), (16). The maximum curtailable power is distributed between [50, 80] kW. The resource parameters are summarized in Appendix A (Tables A1–A3). The distributions of the parameters are depicted in Figure 4. By dividing the load dataset into the training set and the testing set, we train a Long Short-Term Memory (LSTM) model to realize load prediction. The MAPE of the LSTM model on the test set is 7.7% [42,43]. Based on the trained model, we obtained the predicted load. The predicted load and demand response requirement are presented in Figure 5. The standard deviations of the predicted load and demand response requirement are 0.1 and 0.06 times the forecast values, respectively. The distributionally robust risk parameter α is set as 10%. The cost of demand response compensation and the price of natural gas are set as 2.4 RMB/kWh and 0.6 RMB/km^3, respectively. The Time-of-Use (ToU) tariff is illustrated in Figure 6.

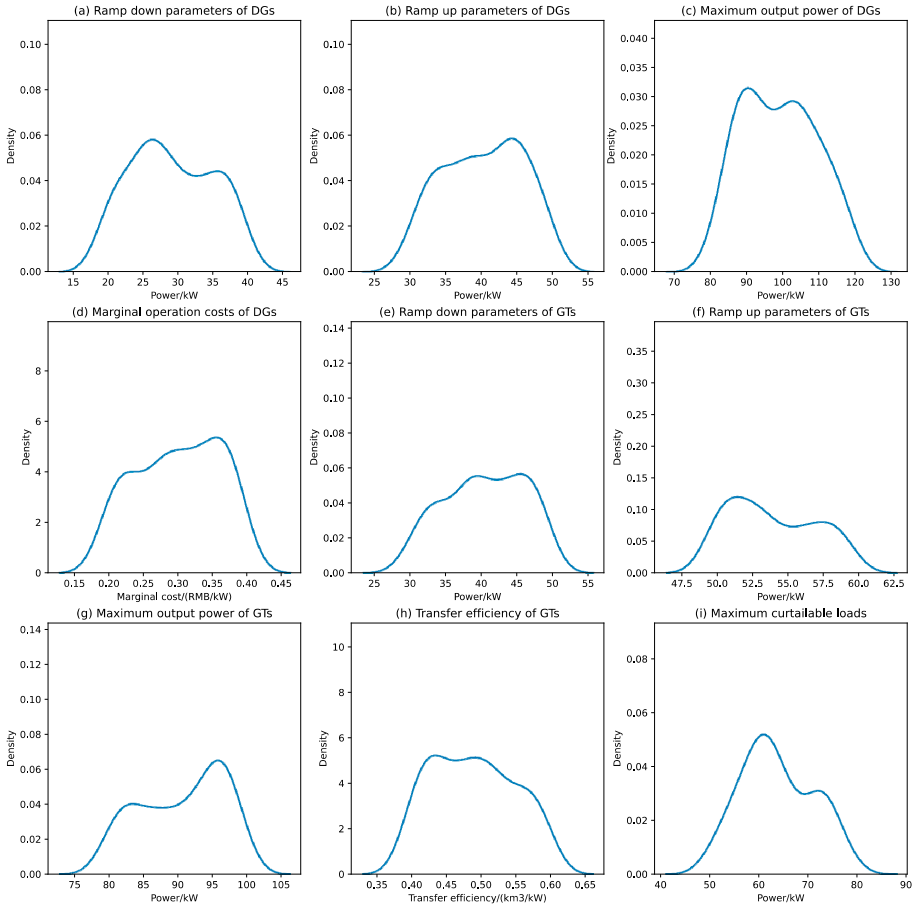

Figure 4. The distributions of DGs, GTs, and curtailable load parameters.

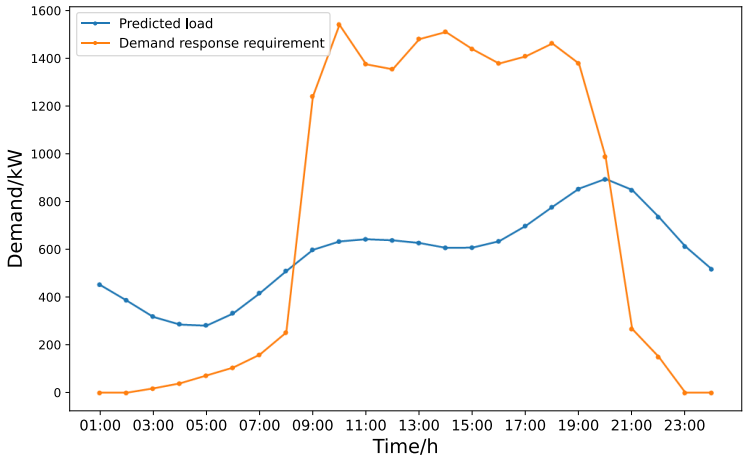

Figure 5. The estimated predicted load and demand response requirement.

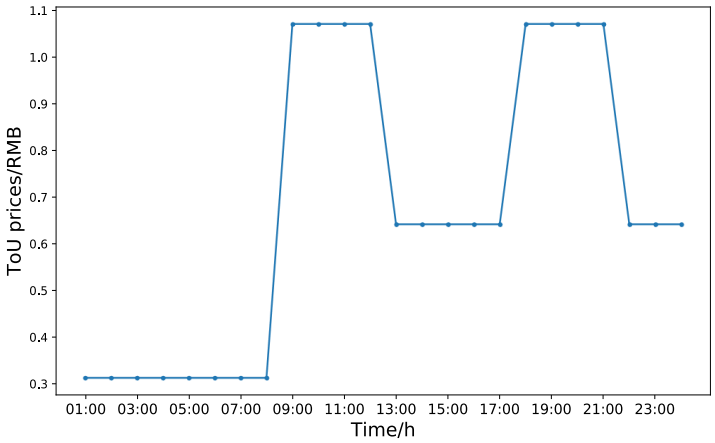

Figure 6. The curve of the ToU tariff.

The architecture of DQN consists of two hidden layers, one input layer, and one output layer. The hidden layers are the fully connected layers using ReLU as the activation function. We observe that the two hidden layers are sufficient to cope with the resource selection. The details of the DQN parameters are listed in Table 2. The model is trained using an Adam optimizer with a learning rate of 1×10^{-4}. The software platform is based on Python combined with Pytorch [44] and Gurobi.

Table 2. Summary of the DQN's parameters.

Item	Value
Batch size	128
No. of hidden layers	2
No. of neurons in the input layer	48
No. of neurons in the output layer	250
No. of neurons in the first hidden layer	256
No. of neurons in the second hidden layer	512
Optimizer	Adam
Learning rate	1×10^{-4}

4.2. Approaches Comparison

This section aims to demonstrate that the proposed approach can find the optimal solution to the problem. Considering the case that all distributed resources in the regional energy network are taking part in the demand response service ($K = 250$), the comparison

candidate has the same formulation as the problem in (26), with the difference being that B_i^{dg}, B_i^{gt}, B_i^c are settled as constants equal to 1. The comparison candidate is solved by Gurobi. Since this process can obtain the global optimal solution, it is regarded as a benchmark. The costs defined in (13) of the two approaches are shown in Tables 3 and 4. The negative value of the cost indicates that the distributed resources can gain a profit by participating in the demand response service. The proposed approach and the comparison candidate have the same profit, indicating that the proposed approach can obtain the optimal solution even though it solves the resource selection problem and power dispatch problem separately.

Table 3. The cost comparison and the number of non-participating resources under 50–150 selected resources.

	K = 50	K = 75	K = 100	K = 125	K = 150
The cost of the proposed approach	−11,859.38	−20,146.53	−23,437.17	−24,486.76	−25,228.14
The cost of the comparison benchmark	\	\	\	\	\
The number of non-participating resources	0	14	29	57	72

Table 4. The cost comparison and the number of non-participating resources under 175–250 selected resources.

	K = 175	K = 200	K = 225	K = 250
The cost of the proposed approach	−25,855.20	−25,849.38	−26,056.01	−26,064.28
The cost of the comparison benchmark	\	\	\	−26,064.28
The number of non-participating resources	94	117	139	180

Furthermore, the results of the power dispatch task are shown in Figure 7, which displays the outputs of DGs and GTs, the load curtailments, and the power purchased from the grid for both the proposed approach and the benchmark comparison. Figure 7 depicts the mean and 95% confidence intervals around the mean for all of the distributed resources. The areas formed by the intervals of the two approaches overlap well, and the mean values are almost identical, indicating that the decision variables obtained by the proposed approach are nearly identical to those of the comparison candidate. This confirms that the proposed approach can achieve the global optimal solution. Additionally, since it is more economical for curtailable loads to participate in the demand response service, they shoulder all of the demand response requirements. Therefore, the outputs of DGs and GTs do not participate in the demand response service. Instead, since the marginal costs of DGs and GTs are relatively cheaper than the ToU price during the flat and peak periods, the outputs of DGs and GTs satisfy all of the predicted load demand in the regional energy network without purchasing power from the grid.

4.3. Results under the Different Number of Selected Resources

In addition to the above discussed case in which $K = 250$, we also consider a scenario in which only part of the distributed resources are chosen to satisfy the predicted load and participate in the demand response service. Accordingly, in this scenario, the value of K is set as a positive integer less than 250. Specifically, we consider cases in which the value of K is equal to 50, 75, 100, 125, 150, 175, 200, and 225. The dynamic training processes of moving average rewards under different values of K are shown in Figure 8. Since the reward is the negative value of the cost, a larger value indicates better performance. During the training process, the agent gradually learns how to maximize the reward, leading to an increase in the moving average reward until convergence. The proposed approach can converge under all scenarios within 100 epochs, which demonstrates the convergence performance.

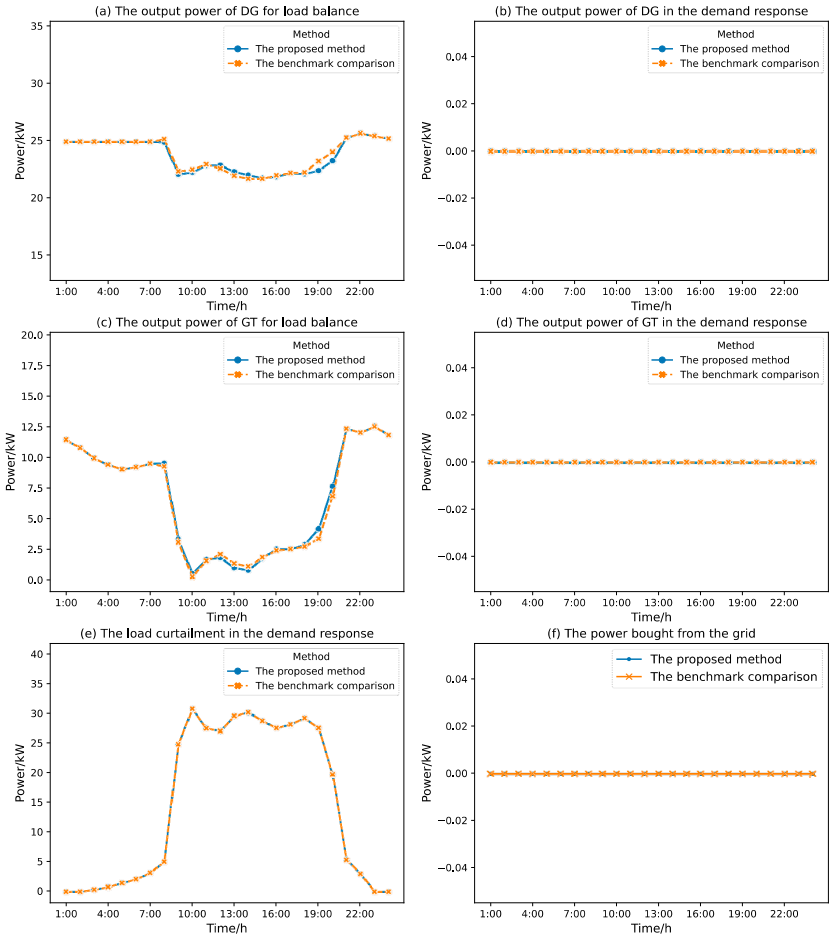

Figure 7. The comparison of decision variables of the proposed and the benchmark approaches.

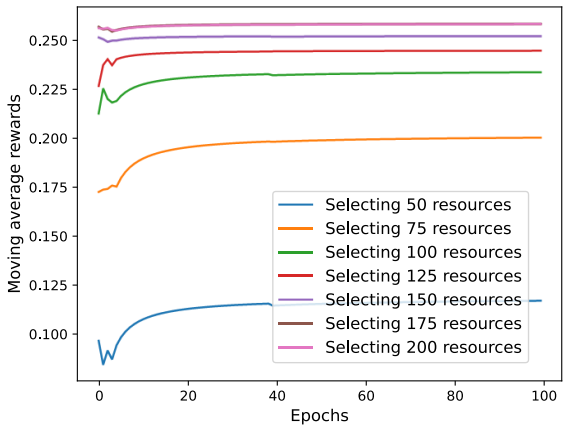

Figure 8. The dynamic training processes under the different number of selected resources.

For different values of K from 50 to 150, Figures 9–12 depict all decision variables' means and 95% confidence intervals around the means. When the value of K equals 50, 75, and 100, DGs and GTs generate power to meet the demand response requirements from 8:00 to 21:00, and curtailable loads contribute the most due to their lower cost. As the number of selected resources increases, the power output of DGs and GTs for demand response decreases. When the value of K is greater than or equal to 150, the demand response task is solely carried out by the curtailable loads. Moreover, with the increase in the value of K, the time slots during which electricity is purchased from the grid decrease. After the value of K increases to 75, the electricity is only bought during the valley period of the ToU tariff.

When the value of K reaches 100, the regional energy system no longer needs to purchase power from the grid. Most of the load demand in the regional energy network is satisfied by the output power of DGs and GTs.

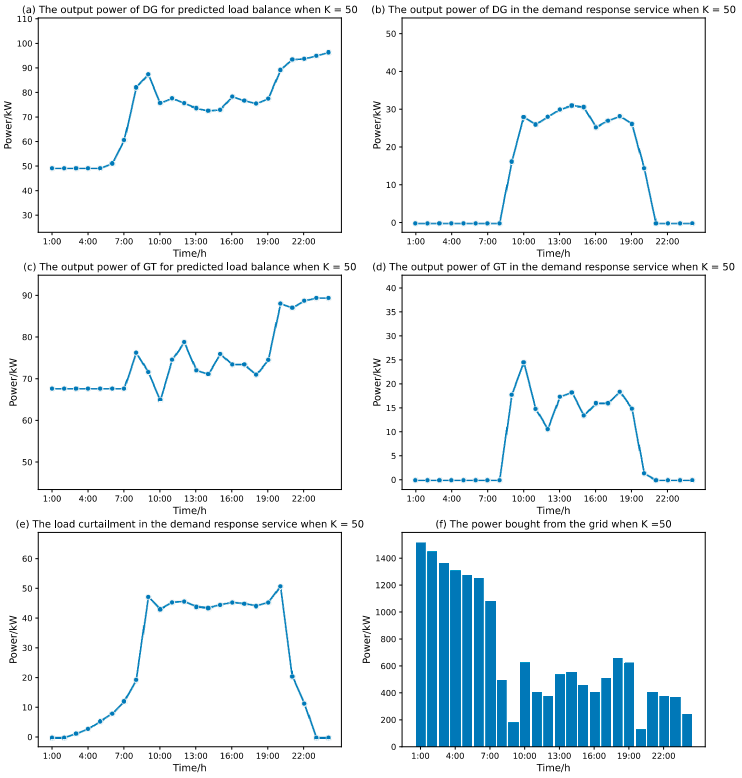

Figure 9. The decision variables of 50 selected resources.

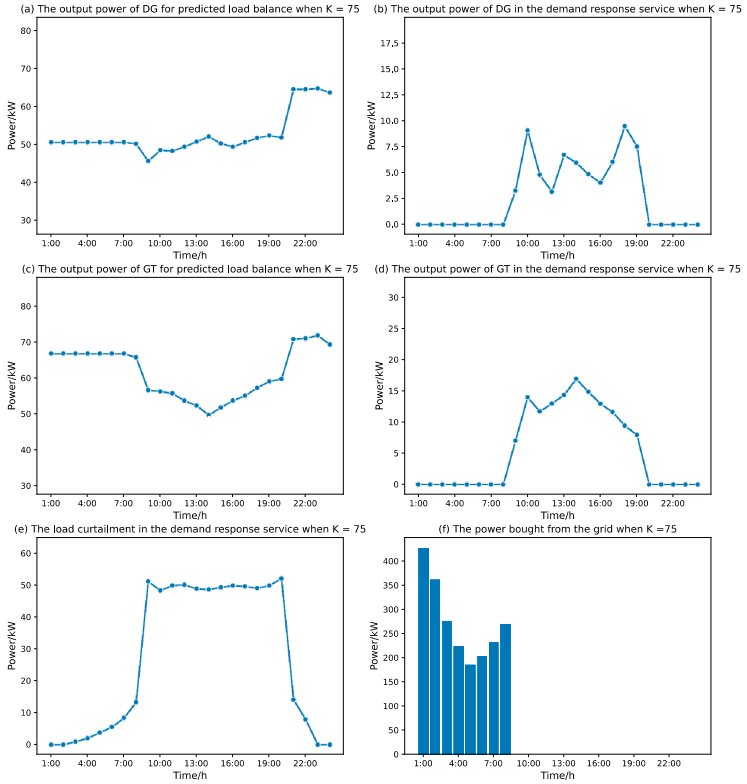

Figure 10. The decision variables of 75 selected resources.

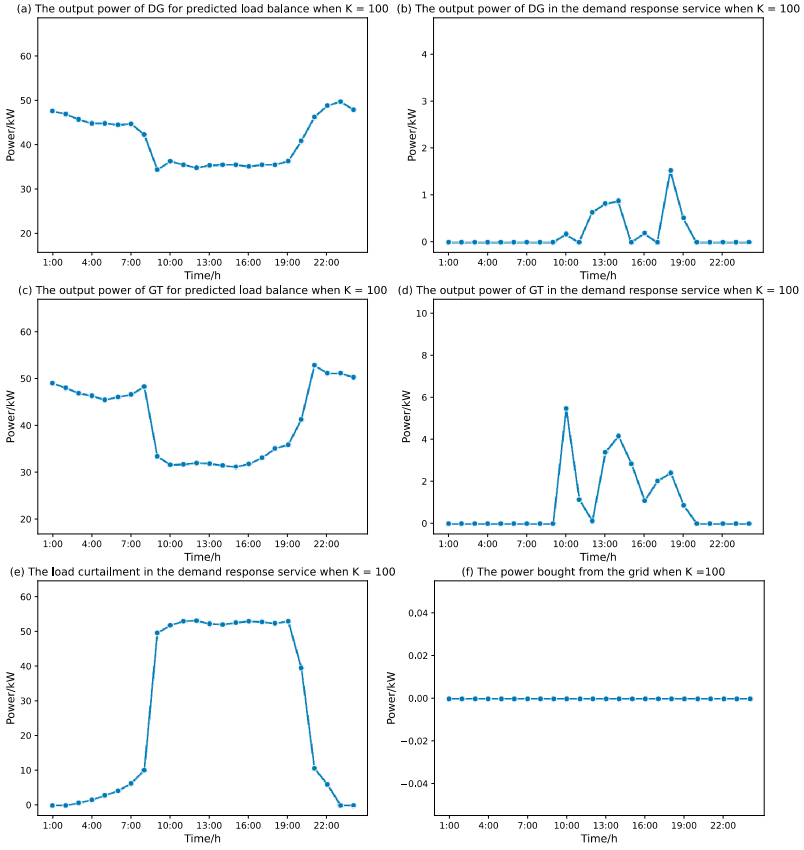

Figure 11. The decision variables of 100 selected resources.

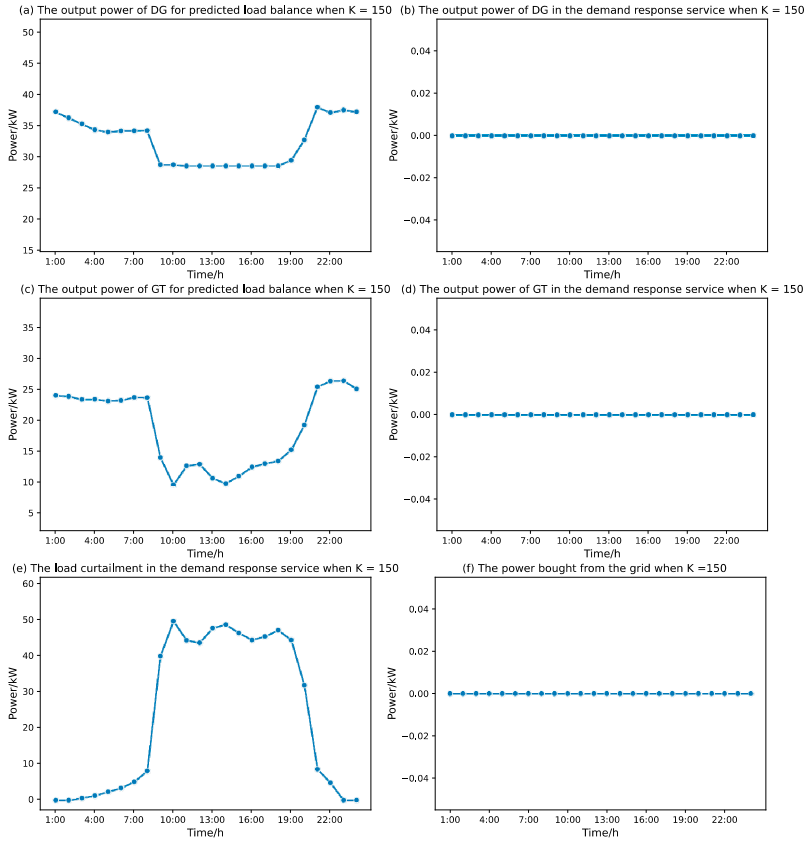

Figure 12. The decision variables of 150 selected resources.

The operation costs defined in (13) of the six scenarios are shown in Tables 3 and 4. The fourth row of Tables 3 and 4 shows the number of distributed resources whose output power or load curtailment is always zero during the scheduled time horizon. We use the term 'non-participating resources' to denote them. Additionally, for comparison, we use the problem in (26) as the comparison candidate, with the difference being that a constraint limiting the number of distributed resources participating in the operation is added:

$$\sum_{i\in\mathcal{A}} B_i^{dg} + \sum_{i\in\mathcal{B}} B_i^{gt} + \sum_{i\in\mathcal{C}} B_i^{c} = K \tag{31}$$

However, the comparison candidate model cannot be solved by commercial solvers due to the limited number of participants; therefore, we use a slash in Tables 3 and 4 to indicate the unsolvable situation. In contrast, the proposed approach is feasible and can still obtain the optimal solution. This demonstrates the superiority of the proposed approach, which decouples the resource selection problem and power dispatch problem so that the computational scale is reduced and the binary and continuous variables realize the decoupling. Thus, the proposed approach allows the regional operator to choose an appropriate number of participants instead of requiring all resources to take part in the demand response service. Additionally, the proposed approach is compared with a multiple attribute decision making approach [45]. The comparative results, as presented in Table 5, reveal that the DQN-based bandit approach can achieve a cost improvement exceeding 4.73% across various K values. The results demonstrate the performance of the proposed approach in identifying optimal solutions for resource selection.

Table 5. The costs improvement of the proposed approach compared to the counterpart.

	K = 75	K = 100	K = 125	K = 150	K = 175	K = 200	K = 225
Improvement	34.38%	15.56%	10.92%	5.99%	7.99%	7.11%	4.73%

Moreover, as depicted in Tables 3 and 4, as more distributed resources participate in the operation, the gained profit increases but the number of non-participating resources also increases, which dampens the resources' initiative to participate in the load balance and demand response service. Therefore, it is essential for the operator to select the proper value of to balance the gained profit and the number of non-participating resources. Moreover, as shown in Table 3, the alteration in cost resulting from the different K values is relatively small for values of K exceeding 100; however, when the value of K falls below 100, the cost undergoes significant changes. This phenomenon can be attributed to the limited capacity of distributed energy resources, which restricts the total capacity of units when excessively small K values are settled. Consequently, a substantial quantity of electricity must be procured from the grid at a considerable expense to fulfill the demand. Conversely, when the K value is large, resource selection competence enables attaining autonomy. Therefore, establishing an appropriate value for K is imperative to ensure the economic viability of the decision.

To determine the proper value of K, we first normalize the values of the profit and the number of non-participating resources. The normalized profit and the number of non-participating resources under different values of K are shown in Figure 13. The normalized profit only increases a little when the value of K equals 100. In contrast, the number of non-participating resources increases a lot at that point. Therefore, the value of K is set as 100 in the following analysis.

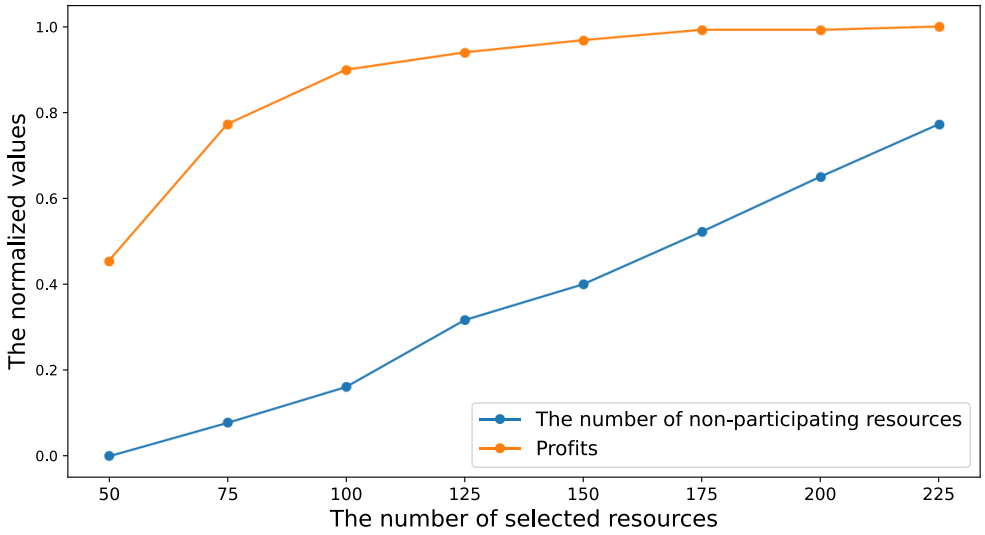

Figure 13. The curves of normalized values of the numbers and profits.

4.4. Comparison with Other Approaches for Dealing with the Uncertainty

Firstly, we evaluate the reliability of the model by simulating different scenarios of the predicted load and demand response requirement. A total of 10,000 samples are generated from the Gaussian distribution for the predicted load and demand response requirement, respectively. We check the violation of the reserve capacity constraints under different distributionally robust risk levels.

The CCP and RO approaches are used for comparison. The CCP approach assumes that the estimation error follows a Gaussian distribution. Then, according to the reference [46], we transform (19) and (20) into the deterministic equivalents:

$$\sum_{i\in\mathcal{A}} \overline{P}_i^{dg}\cdot B_i^{dg} - \sum_{i\in\mathcal{A}} P_{i,t}^{dg}\cdot B_i^{dg} + \sum_{i\in\mathcal{B}} \overline{P}_i^{gt}\cdot B_i^{gt} - \sum_{i\in\mathcal{B}} P_{i,t}^{gt}\cdot B_i^{gt} \geq z_{1-\alpha/2}\cdot\sigma_t^{load} \tag{32}$$

$$\sum_{i\in\mathcal{C}} \overline{P}_i^{cu}\cdot B_i^{c} - \sum_{i\in\mathcal{C}} P_{i,t}^{cu}\cdot B_i^{c} \geq z_{1-\alpha/2}\cdot\sigma_t^{DR} \tag{33}$$

where $z_{1-\alpha/2}$ is the $100\cdot(1-\alpha/2)$th quantile of the standard normal distribution. $\sigma_t^{load}, \sigma_t^{DR}$ are the standard deviations of predicted load and demand response requirement estimation errors. Then, the power dispatch problem can be formulated as follows:

$$\begin{aligned} &\min\ c = \sum_{t\in T} P_t^{grid}\cdot\pi_t + \sum_{i\in\mathcal{A}} c_{dg} + \sum_{i\in\mathcal{B}} c_{gt} + \sum_{i\in\mathcal{C}} c_c \\ &s.t. (2)\text{–}(12), (14)\text{–}(18), (32), (33) \end{aligned} \tag{34}$$

The RO optimization problem to determine the power dispatch in the worst case is formulated as a min-max problem. The objective function is:

$$\min_{x} \max_{\substack{P_t^{load,P}\in\mathcal{Y}_t^{load},\\ P_t^{DR}\in\mathcal{Y}_t^{DR}}} \sum_{t\in T} P_t^{grid}\cdot\pi_t + \sum_{i\in\mathcal{A}} c_{dg} + \sum_{i\in\mathcal{B}} c_{gt} + \sum_{i\in\mathcal{C}} c_c \tag{35}$$

where x is the decision variable, including $P_t^{grid}, P_t^{dg}, P_{i,t}^{dg,DR}, P_t^{gt}, P_{i,t}^{gt,DR}, P_{i,t}^{cu}, B^{dg}, B^{gt}, B_i^c$. $\mathcal{Y}_t^{load}, \mathcal{Y}_t^{DR}$ are the uncertainty set of $P_t^{load,P}$ and P_t^{DR}, respectively. By calculating the quartiles of the history data, we set the uncertainty sets as [1st quartile, 3rd quartile]. It is evident that the optimization solution to the inner maximization problem belongs to

the region where $P_t^{load,P}$ and P_t^{DR} reach their upper limits. Therefore, the problem can be transferred to a deterministic problem as follows.

$$\begin{aligned} &\min c = \sum_{t\in T} P_t^{grid}\cdot\pi_t + \sum_{i\in\mathcal{A}} c_{dg} + \sum_{i\in\mathcal{B}} c_{gt} + \sum_{i\in\mathcal{C}} c_c \\ &s.t.(2)\text{–}(12), (14)\text{–}(16), (37), (38) \end{aligned} \tag{36}$$

$$P_t^{grid} + \sum_{i\in\mathcal{A}} P_{i,t}^{dg}\cdot B_i^{dg} + \sum_{i\in\mathcal{B}} P_{i,t}^{gt}\cdot B_i^{gt} = \hat{P}_t^{+,load,P} + \sum_{i\in\mathcal{C}} \overline{P}_i^{cu} - \sum_{i\in\mathcal{C}} P_{i,t}^{cu}\cdot B_i^{c} \tag{37}$$

$$\sum_{i\in\mathcal{A}} P_{i,t}^{dg,DR}\cdot B_i^{dg} + \sum_{i\in\mathcal{B}} P_{i,t}^{gt,DR}\cdot B_i^{gt} + \sum_{i\in\mathcal{C}} P_{i,t}^{cu}\cdot B_i^{c} = \hat{P}_t^{+,DR} \tag{38}$$

where $\hat{P}_t^{+,load,P}$ and $\hat{P}_t^{+,DR}$ are the upper limits of $\mathcal{Y}_t^{load}, \mathcal{Y}_t^{DR}$, respectively.

We set the risk levels of DRO and CCP to range from 10% to 50% and calculate the possibility of non-violation and profit for the current method and its counterparts. The results are presented in Table 6.

Table 6. Results of non-violation possibility and profit.

Approach	Non-Violation Probability	Profit
DRO-10%	99.94%	23,437.17
CCP-10%	97.90%	23,383.58
DRO-20%	99.66%	23,860.45
CCP-20%	95.38%	23,151.89
DRO-30%	97.11%	23,260.54
CCP-30%	93.46%	23,246.75
DRO-40%	95.31%	23,742.61
CCP-40%	90.78%	23,558.23
DRO-50%	96.84%	24,331.25
CCP-50%	89.26%	22,911.03
RO	100%	6274.31

The non-violation probability comparisons between the current approach and the CCP approach are depicted in Figure 14. The proposed model demonstrates high robustness, with a small chance of violation (greater than 95% chance of non-violation in all scenarios). Of particular importance, when the risk parameter is set to less than 20%, the non-violation probability is greater than 99%. Notably, the proposed approach's probability of violation curve remained above the curve of the CCP model, indicating its higher robustness compared to the CCP approach. Additionally, the RO approach obtained conservative results with no chance of violation.

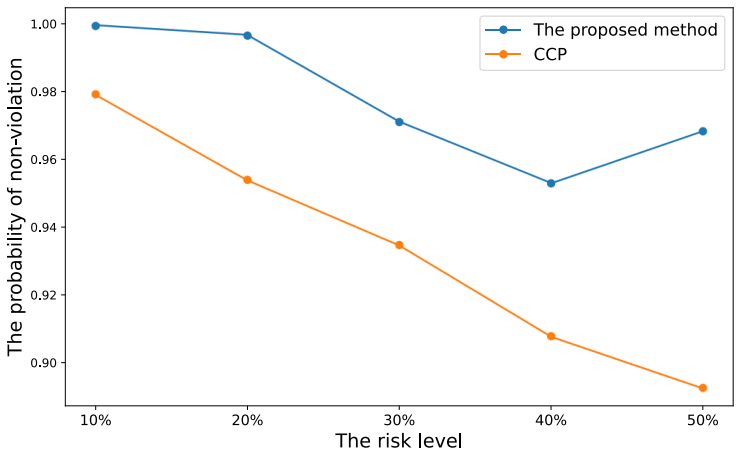

Figure 14. The probability of non-violation between the current approach and the CCP approach.

The difference in profits obtained by the two approaches is presented in Figure 15. The gained profit of the current DRO model is higher than the CCP approach at different risk levels. Specifically, at a risk level of 50%, the relative difference is 6.20%, indicating a significant increase in profits of the current DRO approach compared with the CCP approach. Moreover, the profit of the RO approach is small due to its conservativeness. Hence, the current DRO approach can maintain a high probability of non-violation while reducing the conservativeness to obtain higher profit.

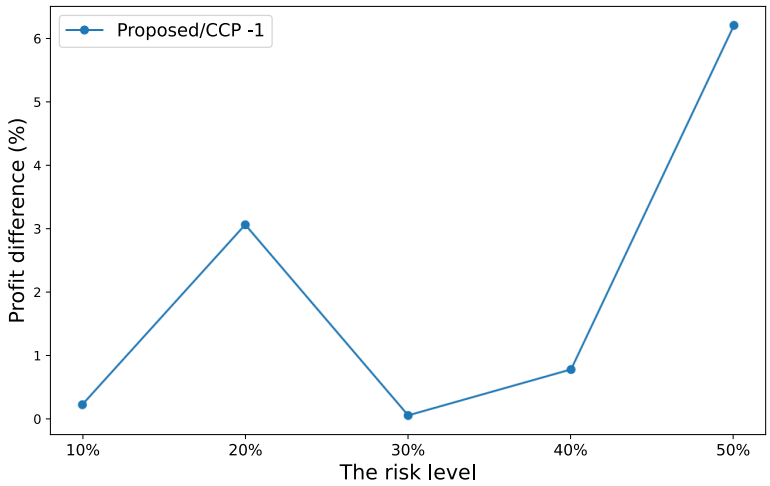

Figure 15. The percentage of profit difference between the proposed approach and the CCP model.

4.5. Comparison with Other Contextual Bandit-Based Approaches

In this subsection, we compare the current approach based on the deep Q-network with the counterpart based on policy gradient (PG)-based approach. A detailed model of the policy gradient (PG)-based bandit approach and the corresponding parameters are provided in Appendix B. A comparison of the profits and non-participating resources of the two approaches is presented in Tables 7 and 8. The comparison curves are depicted in Figure 16.

Table 7. Profits under 50–150 selected resources.

Approach	The DQN-Based Bandit Approach	The PG-Based Bandit Approach	Profit Improvement
$K = 50$	11,859.38	8097.61	46.46%
$K = 75$	20,146.53	17,977.75	12.06%
$K = 100$	23,437.17	20,578.44	13.89%
$K = 125$	24,486.76	24,110.70	1.56%
$K = 150$	25,228.14	25,140.37	0.35%
$K = 175$	25,855.2	25,491.63	1.43%
$K = 200$	25,849.38	25,587.45	1.02%
$K = 225$	26,056.01	25,936.18	0.46%

Table 8. The number of non-participating resources under 50–150 selected resources.

Approach	The DQN-Based Bandit Approach	The PG-Based Bandit Approach
$K = 50$	0	0
$K = 75$	14	6
$K = 100$	29	20
$K = 125$	57	55
$K = 150$	72	74
$K = 175$	94	98
$K = 200$	117	115
$K = 225$	139	139

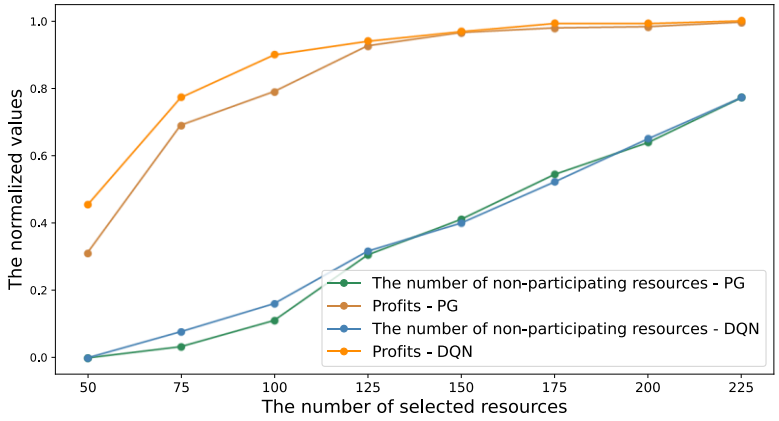

Figure 16. The comparison curves of the DQN-based approach and the PG-based approach.

Based on the results, the profits obtained by the DQN-based approach are higher than the PG-based approach under a different number of selected resources, which indicates the effectiveness of the current DQN-based approach. Notably, even though the number of non-participating resources of the DQN-based approach is greater than that of the counterpart when the number of selected resources is less than 125, the profits also have a noticeable improvement.

5. Conclusions

In this work, we proposed a novel resource management approach to select the proper subset of resources to participate in a demand response service. Resource selection and power dispatch are treated as independent but interrelated tasks. The resource selection policy is learned by a DQN-based agent, with the power dispatch serving as the environment to provide feedback and guide policy learning. Uncertainty in the power dispatch task is addressed by employing a distributionally robust variant.

Case studies validate that the proposed approach can find the optimal solution compared with the benchmark approach, and has good convergence performance. Moreover, both the gained profit and the number of non-participating resources increase along with the increasing number of participants. Furthermore, with the advantage that no assumption on the probability distribution of uncertainty is needed, DRO is applied to cope with the uncertainty in the power dispatch problem. Numerical studies demonstrate the DRO model obtains higher non-violation probability in all scenarios with risk parameters set from 10% to 50% compared to the CCP approach. When the risk parameter is set below 20%, the non-violation probability of DRO can reach more than 99%, which is close to the results of the RO approach, along with a substantial increase in profits compared to the conservative strategy of RO. Moreover, the DQN-based contextual bandit approach can achieve a profit improvement of 0.35–46.46% compared to contextual bandit with policy gradient under different resource selection amounts.

Our proposed approach can tackle the problem which has two intercorrelated decision-making tasks. It will be interesting to apply this approach to other similar problems in the future.

Author Contributions: Conceptualization, Z.L.; Formal analysis, Z.L.; Funding acquisition, Q.A.; Methodology, Z.L.; Software, Z.L.; Supervision, Q.A.; Validation, Z.L. and Q.A.; Writing—original draft, Z.L.; Writing—review and editing, Z.L. and Q.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China, grant number No. 2021YFB2401204.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Appendix A Parameters of Resources in the Regional Energy Network

Table A1. Parameters of Diesel generators.

Item	Value
Resource quantity	100
Marginal costs	[0.2, 0.4]
Ramp up parameters	[30, 50]
Ramp down parameters	[20, 40]
Maximum output	[80, 120]

Table A2. Parameters of Gas turbines.

Item	Value
Resource quantity.	100
Generation efficiencies	[0.4, 0.6]
Ramp up parameters	[50, 60]
Ramp down parameters	[30, 50]
Maximum output	[80, 100]

Table A3. Parameters of Curtailable loads.

Item	Value
Resource quantity	50
Maximum curtailable power	[50, 80]

Appendix B The Detailed Model of the Policy Gradient (PG)—Based Bandit Approach and the Corresponding Parameters

In the policy gradient (PG) -based bandit approach, the raw outputs of the DNN's last layer are denoted by the set $\{x_i\}_{i=1}^{|\mathcal{A}|+|\mathcal{B}|+|\mathcal{C}|}$ with the size $|\mathcal{A}|+|\mathcal{B}|+|\mathcal{C}|$, where x_i is the output of the i-th neuron. Since the possibility of each distributed resource being selected is independent, given the neurons from the output layer $\{x_i\}_{i=1}^{|\mathcal{A}|+|\mathcal{B}|+|\mathcal{C}|}$, the sigmoid activation function is implemented to transform the raw output values of the feedforward neural network into independent probabilities:

$$p(x_i) = \frac{1}{1+e^{-x_i}} \tag{A1}$$

where $p(x_i)$ estimates the possibility of choosing a specific resource i given the input feature.

The decaying ϵ-greedy algorithm in (30) is also used to determine the action in the PG-based approach. If a random number between 0.0 and 1.0 is larger than ε, the top K elements with the larger possibilities are chosen, and their corresponding elements in the action vector are assigned as one. If the random number is smaller than ε, K distributed resources are randomly selected. Therefore, we have $\sum_{i=1}^{|\mathcal{A}|+|\mathcal{B}|+|\mathcal{C}|} a_i = K$.

In the policy gradient approach, DNN learns the policy to maximize the reward:

$$R(\boldsymbol{a})\cdot P(\boldsymbol{a}|\boldsymbol{s},\boldsymbol{\theta}) \tag{A2}$$

where $\boldsymbol{\theta}$ is the parameter of DNN. $\boldsymbol{s}$ is the input context and $\boldsymbol{a} = \{a_i\}_{i=1}^{|\mathcal{A}|+|\mathcal{B}|+|\mathcal{C}|}$ is the output action under the context. The reward $R(\boldsymbol{a})$ under the action $\boldsymbol{a}$ can reflect the quality

of the decided action, which then affects DNN's parameters update. Thus, the unbiased gradient estimation using (30) yielding:

$$\nabla \boldsymbol{\theta} = R(\boldsymbol{a}) \cdot \nabla \log P(\boldsymbol{a}|\boldsymbol{s}, \boldsymbol{\theta}) \tag{A3}$$

Therefore, when the value of $R(\boldsymbol{a})$ is large, the parameter $\boldsymbol{\theta}$ is tuned to increase the possibility of the action $\boldsymbol{a}$. The parameters of DNN are updated by:

$$\boldsymbol{\theta} + \eta \cdot \nabla \boldsymbol{\theta} \rightarrow \boldsymbol{\theta} \tag{A4}$$

where η is the learning rate.

The architecture of DNN consists of two hidden layers, one input layer, and one output layer. The hidden layers are the fully connected layers using tanh as the activation function. Details of the DNN parameters are listed in Table A4. The model is trained using stochastic gradient descent (SGD) optimizer.

Table A4. Summary of the DNN's parameters.

Item	Value
Iteration epochs	1000
No. of neurons in each layer	64
No. of hidden layers	2
No. of neurons in the input layer	48
No. of neurons in the output layer	250
Optimizer	SGD
Learning rate	1×10^{-3}

References

1. Available online: http://www.nea.gov.cn/ (accessed on 2 June 2022).
2. Available online: https://www.gov.cn/ (accessed on 2 April 2021).
3. Zhang, Y.; Ai, Q.; Wang, H.; Li, Z.; Huang, K. Bi-level distributed day-ahead schedule for islanded multi-microgrids in a carbon trading market. *Electr. Power Syst. Res.* **2020**, *186*, 106412. [CrossRef]
4. Haider, R.; Annaswamy, A.M. A hybrid architecture for volt-var control in active distribution grids. *Appl. Energy* **2022**, *312*, 118735. [CrossRef]
5. Yang, X.; Wang, Z.; Zhang, H.; Ma, N.; Yang, N.; Liu, H.; Zhang, H.; Yang, L. A Review: Machine Learning for Combinatorial Optimization Problems in Energy Areas. *Algorithms* **2022**, *15*, 205. [CrossRef]
6. Bertsimas, D.; Litvinov, E.; Sun, X.A.; Zhao, J.; Zheng, T. Adaptive Robust Optimization for the Security Constrained Unit Commitment Problem. *IEEE Trans. Power Syst.* **2013**, *28*, 52–63. [CrossRef]
7. Kong, F.; Liu, Y.; Tong, L.; Guo, W.; Qiu, Y.; Wang, Y. Optimization of co-production air separation unit based on MILP under multi-product deterministic demand. *Appl. Energy* **2022**, *325*, 119850. [CrossRef]
8. Khodayar, M.; Liu, G.; Wang, J.; Khodayar, M.E. Deep learning in power systems research: A review. *CSEE J. Power Energy Syst.* **2020**, *7*, 209–220.
9. Matsuo, Y.; Le Cun, Y.; Sahani, M.; Precup, D.; Silver, D.; Sugiyama, M.; Uchibe, E.; Morimoto, J. Deep learning, reinforcement learning, and world models. *Neural Netw.* **2022**, *152*, 267–275. [CrossRef]
10. Jiang, W.; Liu, Y.; Fang, G.; Ding, Z. Research on short-term optimal scheduling of hydro-wind-solar multi-energy power system based on deep reinforcement learning. *J. Clean. Prod.* **2023**, *385*, 135704. [CrossRef]
11. Li, Z.; Sun, Z.; Meng, Q.; Wang, Y.; Li, Y. Reinforcement learning of room temperature set-point of thermal storage air-conditioning system with demand response. *Energy Build.* **2022**, *259*, 111903. [CrossRef]
12. Qiu, D.; Ye, Y.; Papadaskalopoulos, D.; Strbac, G. A deep reinforcement learning method for pricing electric vehicles with discrete charging levels. *IEEE Trans. Ind. Appl.* **2020**, *56*, 5901–5912. [CrossRef]
13. Ye, Y.; Qiu, D.; Sun, M.; Papadaskalopoulos, D.; Strbac, G. Deep reinforcement learning for strategic bidding in electricity markets. *IEEE Trans. Smart Grid* **2020**, *11*, 1343–1355. [CrossRef]
14. Du, Y.; Zandi, H.; Kotevska, O.; Kurte, K.; Munk, J.; Amasyali, K.; Mckee, E.; Li, F. Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. *Appl. Energy* **2021**, *281*, 116117. [CrossRef]
15. Bouneffouf, D.; Rish, I.; Cecchi, G.A.; Feraud, R. Context attentive bandits: Contextual bandit with restricted context. *arXiv* **2017**, arXiv:1705.03821.
16. Zhang, Y.; Wu, Q.; Ai, Q.; Catalão, J.P.S. Closed loop Aggregated Baseline Load Estimation using Contextual Bandit with Policy Gradient. *IEEE Trans. Smart Grid* **2022**, *13*, 243–254. [CrossRef]

17. Audibert, J.Y.; Munos, R.; Szepesvári, C. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.* **2009**, *410*, 1876–1902. [CrossRef]
18. Silva, N.; Werneck, H.; Silva, T.; Pereira, A.C.M.; Rocha, L. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Syst. Appl.* **2022**, *197*, 116669. [CrossRef]
19. Lu, S.; Zhou, Y.H.; Shi, J.C.; Zhu, W.; Yu, Q.; Chen, Q.G.; Da, Q.; Zhang, L. Non-stationary Continuum-armed Bandits for Online Hyperparameter Optimization. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event, AZ, USA, 21–25 February 2022; pp. 618–627.
20. Kim, G.S.; Hong, Y.S.; Lee, T.H.; Paik, M.C.; Kim, H. Bandit-supported care planning for older people with complex health and care needs. *arXiv* **2023**, arXiv:2303.07053.
21. Yang, Y.; Teng, X.; Chen, K.; Cheng, W.; Si, Y.; Xu, J. Multi-armed Bandit Based Load Aggregation for Power System Frequency Regulation. In Proceedings of the 16th Annual Conference of China Electrotechnical Society, Beijing, China, 25–26 September 2021; Springer Nature Singapore: Singapore, 2022; Volume 2, pp. 686–693.
22. Chen, X.; Hu, Q.; Shi, Q.; Quan, X.; Wu, Z.; Li, F. Residential HVAC aggregation based on risk-averse multi-armed bandit learning for secondary frequency regulation. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 1160–1167. [CrossRef]
23. Lei, Z.; Xu, X.; Li, J.; Fan, L.; Chen, X.; Ding, H. Optimal scheduling of Renewable Energy Sources for Grid Frequency Stability Using Multi-armed Bandit Method. In Proceedings of the 2021 IEEE Sustainable Power and Energy Conference (iSPEC), Nanjing, China, 23–25 December 2021; pp. 665–670.
24. Sun, J.; Zhao, Y.; Zhang, N.; Chen, X.; Hu, Q.; Song, J. A dynamic distributed energy storage control strategy for providing primary frequency regulation using multi-armed bandits method. *IET Gener. Transm. Distrib.* **2022**, *16*, 669–679. [CrossRef]
25. Hu, Q.; Zhang, N.; Quan, X.; Bai, L.; Wang, Q.; Chen, X. A user selection algorithm for aggregating electric vehicle demands based on a multi-armed bandit approach. *IET Energy Syst. Integr.* **2021**, *3*, 295–305. [CrossRef]
26. Cheng, S.; Han, R.; Zhao, Y.; Hu, Q.; Jiang, W. Aggregating residential demands with a multi-armed bandit approach. In Proceedings of the 2019 IEEE Sustainable Power and Energy Conference (iSPEC), Beijing, China, 21–23 November 2019; pp. 2144–2149.
27. Chen, X.; Nie, Y.; Li, N. Online residential demand response via contextual multi-armed bandits. *IEEE Control. Syst. Lett.* **2020**, *5*, 433–438. [CrossRef]
28. Zhao, J.; Zhang, M.; Yu, H.; Ji, H.; Song, G.; Li, P.; Wang, C.; Wu, J. An islanding partition method of active distribution networks based on chance-constrained programming. *Appl. Energy* **2019**, *242*, 78–91. [CrossRef]
29. Doluweera, G.; Hahn, F.; Bergerson, J.; Pruckner, M. A scenario-based study on the impacts of electric vehicles on energy consumption and sustainability in Alberta. *Appl. Energy* **2020**, *268*, 114961. [CrossRef]
30. Ehsan, A.; Yang, Q. Scenario-based investment planning of isolated multi-energy microgrids considering electricity, heating and cooling demand. *Appl. Energy* **2019**, *235*, 1277–1288. [CrossRef]
31. Li, Y.; Ming, B.; Huang, Q.; Wang, Y.; Liu, P.; Guo, P. Identifying effective operating rules for large hydro–solar–wind hybrid systems based on an implicit stochastic optimization framework. *Energy* **2022**, *245*, 123260. [CrossRef]
32. Ben-Tal, A.; El Ghaoui, L.; Nemirovski, A. *Robust Optimization*; Princeton University Press: Princeton, NJ, USA, 2009.
33. Xie, W.; Ahmed, S. Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation. *IEEE Trans. Power Syst.* **2017**, *33*, 1860–1867. [CrossRef]
34. Jin, X.; Liu, B.; Liao, S.; Cheng, C.; Yan, Z. A Wasserstein metric-based distributionally robust optimization approach for reliable-economic equilibrium operation of hydro-wind-solar energy systems. *Renew. Energy* **2022**, *196*, 204–219. [CrossRef]
35. Erdoğan, E.; Iyengar, G. Ambiguous chance constrained problems and robust optimization. *Math. Program.* **2006**, *107*, 37–61. [CrossRef]
36. Amari, S.I.; Karakida, R.; Oizumi, M. Information geometry connecting wasserstein distance and kullback–leibler divergence via the entropy-relaxed transportation problem. *Inf. Geom.* **2018**, *1*, 13–37. [CrossRef]
37. Xie, W. On distributionally robust chance constrained programs with asserstein distance. *Math. Program.* **2021**, *186*, 115–155. [CrossRef]
38. Bayraksan, G.; Love, D.K. Data-driven stochastic programming using phi-divergences. *Oper. Res. Revolut. Inf.* **2015**, *10*, 1–19.
39. Hanasusanto, G.A.; Roitch, V.; Kuhn, D.; Wiesemann, W. Ambiguous joint chance constraints under mean and dispersion information. *Oper. Res.* **2017**, *63*, 751–767. [CrossRef]
40. Zhang, Y.; Shen, S.; Mathieu, J. Distributionally robust chance constrained optimal power flow with uncertain renewables and uncertain reserves provided by loads. *IEEE Trans. Power Syst.* **2017**, *32*, 1378–1388. [CrossRef]
41. Schofield, J.; Tindemans, S.; Carmichael, R.; Tindemans, S.H.; Bilton, M.; Woolf, M.; Strbac, G. Low carbon london project: Data from the dynamic time-of-use electricity pricing trial, 2013. *Tech. Rep.* **2016**, *7857*, 7857.
42. Al Mamun, A.; Sohel, M.; Mohammad, N.; Sunny, M.S.H.; Dipta, D.R.; Hossain, E. A comprehensive review of the load forecasting techniques using single and hybrid predictive models. *IEEE Access* **2020**, *8*, 134911–134939. [CrossRef]
43. Feng, Y.; Rios, I.; Ryan, S.M.; Spürkel, K.; Watson, J.P.; Wets, R.J.B.; Woodruff, D.L. Toward scalable stochastic unit commitment. Part 1: Load scenario generation. *Energy Syst.* **2015**, *6*, 309–329. [CrossRef]
44. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703.

45. Hwang, C.L.; Yoon, K. Methods for multiple attribute decision making. In *Multiple Attribute Decision Making: Methods and Applications: A State-of-the-Art Survey*; Springer: New York, NY, USA, 1981; Volume 186, pp. 58–191.
46. Wu, H.; Shahidehpour, M.; Li, Z.; Tian, W. Chance-constrained day-ahead scheduling in stochastic power system operation. *IEEE Trans. Power Syst.* **2014**, *29*, 1583–1591. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.