Duck Curve Aware Dynamic Pricing and Battery Scheduling Strategy Using Reinforcement Learning

Daichi Watari, *Member, IEEE*, Ittetsu Taniguchi, *Senior Member, IEEE*, and Takao Onoye, *Senior Member, IEEE*

***Abstract*—The duck curve is becoming a global problem in energy technology due to the rapid increase in solar power adoption and the rise of prosumers. To address this issue, a resource aggregator (RA) has emerged to provide flexible solutions through aggregating the prosumers and demand response such as dynamic pricing. This paper proposes an optimal strategy for the RA that dispatches dynamic pricing to the prosumers and leverages the battery system at both RA and prosumer levels. The proposed method is based on a model-free deep reinforcement learning (DRL) algorithm to optimize each prosumer's retail prices and schedule usage of the RA's battery power station. An objective reward function is used to maximize the RA's profit, minimize the prosumer's cost, and maximize the improvement of the duck curve. The performance of the proposed DRL-based strategy was demonstrated by simulation experiments using actual wholesale price, demand, and PV generation data. The results show that the proposed strategy can improve the standard deviation and peak-to-average ratio of net load by up to 57.1% and 23%, respectively.**

***Index Terms*—Duck curve, demand response, dynamic pricing, battery scheduling, deep reinforcement learning, prosumer.**

NOMENCLATURE

Indices

t, k	index for time steps, $t \in \{1, 2, \dots, T\}$ and $k \in \{1, 2, \dots, t-1\}$
Δt	time interval
n	index for prosumers, $n \in \{1, 2, \dots, N\}$

Parameters

$\xi_{t,n}$	price elasticity of prosumer n at time t
μ_t	wholesale electricity price at time t
μ_{min}, μ_{max}	minimum, maximum wholesale electricity price of day
κ_t	purchase price for selling excess energy from prosumers at time t
ν	coefficient of price limit for retailed electricity price $\lambda_{t,n}$
$\lambda^{lb}, \lambda^{ub}$	minimum, maximum retailed electricity price
S_n	patience period of prosumer n
α_n, β_n	coefficients of dissatisfaction value $U_{t,n}$ of prosumer n
C_n^{bat}	battery capacity of prosumer n
$C_{ch,n}^{rate,bat}, C_{disch,n}^{rate,bat}$	maximum charging, discharging rate of prosumer n's battery
$\eta_{ch,n}^{bat}, \eta_{disch,n}^{bat}$	charging, discharging efficiencies of prosumer n's battery
$e_n^{lb,bat}, e_n^{ub,bat}$	minimum, maximum energy rate of battery charging/discharging of prosumer n
λ_n^{th}	price threshold for rule-based battery controller of prosumer n
h_n	price-sensitivity coefficient for battery controller of prosumer n
C^{ra}	capacity of RA's battery power station
$C_{ch}^{rate,ra}, C_{disch}^{rate,ra}$	maximum charging, discharging rate of RA's battery power station
$\eta_{ch}^{ra}, \eta_{disch}^{ra}$	charging, discharging efficiencies of RA's battery power station
ω_1, ω_2	weight coefficients for reward function of RA

Variables

$e_{t,n}^{net}$	net load of prosumer n at time t
$e_{t,n}^{pv}$	PV generation of prosumer n at time t
$e_{t,n}^{bat}$	battery power of prosumer n at time t
$e_{t,n}^{dm}$	total load demand of prosumer n at time t
$e_{t,n}^{inelas}$	inelastic load of prosumer n at time t
$e_{t,n}^{elas}$	elastic load of prosumer n at time t
$e_{t,n}^{defer}$	deferred load of prosumer n at time t
$e_{t,n}^{shft}$	shifted load of prosumer n at time t
$\lambda_{t,n}$	retail price for prosumer n at time t
$u_{t,k,n}$	binary decision variable whether curtailed energy of prosumer n at time k is scheduled to time t
$U_{t,n}$	dissatisfaction value of prosumer n at time t
$SOC_{t,n}^{bat}$	SOC level of prosumer n's battery at time t
$C_{t,n}^{pro}$	total electricity cost of prosumer n at time t

Manuscript received 22 April 2022; revised 15 September 2022, 13 February 2023, and 20 May 2023; accepted 18 June 2023. Date of publication 21 June 2023; date of current version 26 December 2023. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP21J10312 and Grant 22H03697. Paper no. TSG-00559-2022. *(Corresponding author: Daichi Watari.)*

The authors are with the Graduate School of Information Science and Technology, Osaka University, Suita 565-0871, Osaka, Japan (e-mail: watari.daichi@ist.osaka-u.ac.jp; i-tanigu@ist.osaka-u.ac.jp; onoye@ist.osaka-u.ac.jp).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2023.3288355.

Digital Object Identifier 10.1109/TSG.2023.3288355

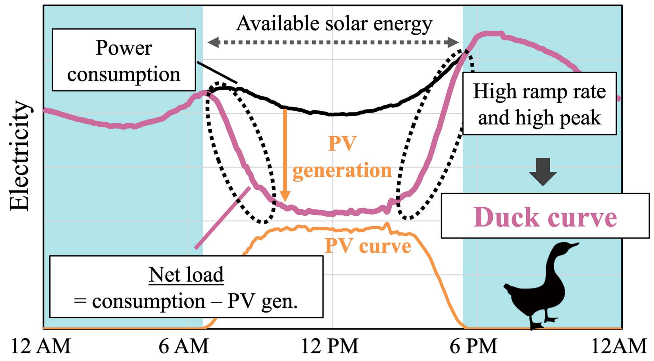


Fig. 1. Example of typical duck curve graph.

E_t^{ra}	battery power of RA's battery power station at time t
SOC_t^{ra}	SOC level of RA's battery power station at time t
E_t^{net}	total net load that RA trades with markets at time t
P_t^{ra}	revenue through electricity retail of RA at time t
R_t^{duck}	remuneration for duck curve improvement at time t
$e_{t,n}^{net,org}$	net load originally scheduled of prosumer n at time t
$e_{t,n}^{dm,org}$	total load demand originally scheduled of prosumer n at time t

I. Introduction

THE HIGH market penetration of installed solar energy systems has turned many consumers into prosumers. Although the increase in prosumers installing photovoltaic (PV) generation systems accelerates the decarbonization of the power grid, it can also cause a severe problem, namely, a "duck curve" [1]. Fig. 1 shows a conceptual diagram of a duck curve. This graph illustrates an example of a net load profile, defined as the total power consumption minus variable renewable energy (PV generation) of an entire grid on a given day. The duck curve is a particular case of a net load graph in a scenario with high PV penetration. The timing imbalance between the demand peak and solar production causes a steep duck curve and a demand peak valley. To cope with this high ramping rate, an independent system operator (ISO) needs to augment delivered power with conventional supply sources such as gas or coal power plants. However, the sudden start-up of traditional sources increases carbon emissions and makes the power grid inefficient and expensive. The duck curve problem is also becoming severe in Japan [2] and other regions with high levels of PV penetration [3], even with the reduced demand in the COVID-19 pandemic period [4].

The conventional solution to the duck curve is the development of supply-side flexibility that includes retrofitting fossil fuel power plants [5], adjusting the orientation of the PV module in solar plants [6], and improving the efficiency of the unit commitment schedule in electric power production [7]. However, these supply-side approaches have limitations in solving the duck curve due to the continuous growth of PV penetration levels in most countries [8]. PV generation is intermittent and non-dispatchable, and its production quantity often changes on an hourly, daily, and seasonal basis. This variability leads to a significant cost paid to develop the flexibility capacity of the grid [9]. Besides, A day-ahead market usually treats the duck curve problem [10], but the duck curve is also becoming a critical issue in real-time. The forecasting of PV generation is generally a difficult task, especially day-ahead forecasting [11]. Therefore, in a power system with a high penetration of PV, the risk of mismatch with day-ahead forecasting (overestimation and underestimation) is not small and cannot be ignored in real-time. On the other hand, demand-side flexibility has become attractive because recent prosumers have control over many types of flexible loads such as schedulable appliances, batteries, and electric vehicles (EVs), to manage intermittent renewable energy [12].

As for the demand-side flexibility, battery systems are promising to reshape the load curve locally [13]. The installation of PV panels on the demand side accounts for a large share of the entire PV capacity, e.g., 30% in the US [14]. The PV on the demand side, such as rooftop and behind-the-meter PV panels, is a major factor of the duck curve, as well as solar PV at the utility-scale. Although the supply side has no access to demand-side PV information, the demand side can locally monitor PV generation and effectively schedule its own battery. To effectively solve the duck curve, the use of demand-side batteries is also important.

One of the efficient solutions to induce demand-side flexibility is the implementation of a demand response (DR) program. DR programs are typically price-based or incentive-based schemes that encourage end-users to change their behaviors. They aim to achieve the desired shape of demand load, e.g., peak-shaving, valley filling, and load shifting [15]. Among DR programs, the price-based DR program has been used in many countries due to its scalability and efficiency. Depending on the degree of dynamic pricing, the price-based DR programs are mainly classified into critical peak pricing (CPP), time of use (TOU), and real-time pricing (RTP) [16]. In particular, the RTP has significant potential to change end-user behavior, as demonstrated in the literature [17].

To implement DR programs, a resource aggregator (RA) plays a critical role in efficiently coordinating the end-user response [18]. The RA is one of the market participants and is responsible for many roles, such as market participation, controlling own energy resources, and DR implementation to prosumers, as an integrator between the market and prosumers [19]. The RA generally aggregates energy resources such as energy storage facilities [20] and the prosumer demand controlled by DR programs and provides ancillary services to the grid. The business model of the RA is to earn revenue through electricity retailing to prosumers and to receive remuneration from the ISO by providing ancillary services, such as improving the duck curve. The power aggregation of many prosumers is suitable for efficient DR programs and achieving demand-side flexibility.

Many studies have been conducted on the implementation effect of DR. Jiang et al. [21] proposed an RTP model based on the matrix of electricity price elasticity that expresses the relationship between retail prices and customer response. Wang et al. [22] developed a fairness-aware RTP mechanism based on an optimization approach and established a residential user evaluation system with indicators for user characteristics. Yang et al. [23] proposed the energy optimization method based on an integrated DR program by a multi-energy provider to achieve a win-win strategy for a utility provider and its customers. Taherian et al. [24] integrated load forecasting and a metaheuristic-based RTP model to maximize the profit of utility providers and minimize the electricity costs of both proactive and reactive customers. The aforementioned studies have investigated the pricing mechanism and implementation scheme, but they mainly focused on demand-side profitability for the RA and customers and lack the pricing scheme to improve the supply-side problem, especially for the duck curve.

Many researchers have focused on improving the duck curve through energy management and demand response coordinated by aggregators. First, we introduce related works that assume cooperative scenarios, i.e., the demand-side load is managed by a central optimizer. In one work [25], the potential of demand-side flexibility, derived from large EV fleets, was investigated to minimize ramp-up requirements. Howlader et al. [26] proposed an optimal thermal unit commitment considering RTP and demand-side load based on mixed-integer linear programming (MILP) to fill peak and off-peak gaps in the duck curve. In another work [27], the feasibility and potential of pre-cooling strategies in residential households were demonstrated to mitigate the duck curve. Yoon et al. [28] formulated a dynamic pricing DR strategy for building heating, ventilation, and air conditioning (HVAC) systems to reduce peak load as a single-level optimization model. It is doubtful that these cooperative scenarios are feasible in reality, since end users may feel discomfort by the privacy issue and having their own devices controlled by someone else.

Next, we introduce related works that assume noncooperative scenarios, i.e., aggregators indirectly controlling end-users through DR programs. Ferdous et al. formulated a nonlinear programming (NLP) problem to perform optimal dynamic pricing by electricity retailers [29]. In Zhang et al. [30], the concept of vehicle-to-grid (V2G) under dynamic pricing used to mitigate the ramp event in the duck curve was formulated as a Stackelberg game. Sheha et al. [10] also proposed a Stackelberg game framework to solve the duck curve. They assumed that households optimize both battery systems and HVAC systems to minimize electricity costs under dynamic pricing. All of these works proposed model-based approaches based on mathematical optimization. However, these approaches made the impractical assumption of having complete knowledge of the end-user systems, and the computational cost is expensive. In addition, they require forecast profiles of solar generation and power consumption during the planning period. Since the model-based approach is deterministic, uncertainties such as forecast error can cause failure to improve the duck curve.

There are several approaches to address various uncertainties in a model-based optimization approach. Stochastic optimization accounts for uncertainties by considering a large number of scenarios [31] or reduced scenarios [32]. However, the computational complexity of stochastic models is generally expensive, and the accurate probability distribution of uncertain variables is necessary, which is a time-consuming process. Robust optimization obtains the solution within certain sets of uncertain variables in the worst-case scenario [33] and the upper / lower bounds [34]. However, the obtained solution tends to be conservative, and the accurate range of uncertain variables needs to be known in advance, which is an unpractical setting. Both stochastic and robust optimization are model-based approaches, which makes it challenging to accurately model the nonlinear and complex customer response to dynamic pricing. Meanwhile, model-free approaches, such as reinforcement learning, can learn from data not only uncertainties but also nonlinear relationships.

We address the above research challenges by employing a model-free reinforcement learning (RL) approach. RL is an area of machine learning that attempts to learn which actions are the best in environments from data without expert knowledge of the system [35]. In particular, deep RL (DRL), which combines deep neural networks with RL, is known to perform well in decision-making for high-dimensional problems, such as power systems [36]. Several studies have proposed an aggregator strategy using RL and DRL. Qiu et al. [37] proposed a DRL approach to determine the charging prices of EVs in the EV aggregator. However, their strategy is tailored to DR programs for charging EVs, and they do not consider prosumers. Lu et al. [38] proposed a model-free price-based DR strategy for the electricity retailer based on Q-learning, a typical RL algorithm. Kuang et al. [39] proposed an optimal incentive-based DR strategy based on DRL for a virtual power plant (VPP) considering the customer's risk attributes. However, these studies did not consider the duck curve problem or prosumers who use renewable energy and did not support the sale of surplus energy back to the aggregator. Therefore, we can conclude that there is no significant study on the duck-curve-improvement RA's strategy, using the RL/DRL algorithm and considering prosumers.

In this study, we propose a model-free DRL-based strategy for RAs to improve the duck curve to address computer complexity, environmental uncertainty, and privacy concerns for prosumers. The RA's actions include dynamic pricing for end-users (prosumers) and the power-use scheduling of a battery power station owned by RA. Consequently, the objective of our method is to maximize the RA's profit, the prosumer's cost-savings, and the improvement of the duck curve. We demonstrate the performance of the proposed method by comparing it with specific baselines and examining the impact of different scenarios and parameters on its performance. To the best of our knowledge, this is the first study aiming to improve the duck curve with real-time RA strategies. The main contributions of this paper are as follows.

- We propose a model-free DRL-based algorithm to make the RA learn the optimal strategy for solving the duck curve. The trained strategy can calculate dynamic pricing and battery scheduling (DP-BS) in real-time without complete knowledge of the prosumers.

- We extend a hierarchical energy market model [38] to include prosumers with a PV panel and a battery system. It is formulated as a Markov Decision Process (MDP) model to apply the DRL algorithm.
- The response of prosumers to retail electricity prices is modeled as two price-responsive devices: elastic load and battery use. This modeling helps mimic a real-world system as a proof-of-concept for the proposed method.
- The design of the reward function is carefully explored to improve the duck curve.
- Simulation experiments are conducted to demonstrate the performance of the proposed strategy from multiple aspects, such as a mitigation of the duck curve, balancing of prosumer cost and RA profit, and the RA's battery size.

This paper extends the work presented in [40]. The earlier work only focused on load curtailment in response to retail prices as prosumer behavior. Additionally, the prosumer's battery controller was too basic, only allowing for the battery to be charged or discharged at maximum rates and not taking into account the retail price levels and the current states of demand and PV generation. In this paper, we present a more comprehensive model of prosumer behavior that introduces the capability of load shifting in the prosumer's demand load instead of load curtailment. Moreover, we enable the prosumer's battery controller to control the battery continuously in proportion to the level of retail prices and to incorporate the current demand and PV balances. These modifications make the system more realistic and able to address the duck curve problem more effectively. The MDP model trained by the DRL algorithm has also been improved by incorporating additional important states, leading to better solutions. To demonstrate the effectiveness of our method, we performed a sensitivity analysis for the battery size and weight parameters in the reward functions, and conducted a full-year simulation beyond the 1-week simulation performed in the previous study.

The organization of this paper is as follows. Section II presents the problem setting and system models, and Section III provides the proposed RA's strategy for improving the duck curve. In Section IV, a simulation is conducted to obtain results to verify the performance of the proposed method. Finally, Section V concludes this paper.

II. Problem Statement

In this section, we present a detailed problem setting and system model. As shown in Fig. 2, our target system is a hierarchical energy market composed of supply-side and demand-side sectors [38]. The recent development of ICT (information and communication technology) enables bi-directional communication of information between these entities in real-time. On the supply side, the wholesale electricity market (WEM) is electricity is traded, sending agreed wholesale electricity prices to the entire system. The ISO is responsible for monitoring the state of the grid and resolving issues such as the duck curve by offering flexibility on the supply and grid side. The demand side consists of RA, a battery power station, and prosumers. We assume that the RA owns a controllable battery power station as an energy storage system and joins the WEM. The RA aggregates the net load by the prosumers and the battery power station and then trades it with the WEM and the ISO. Prosumers have PV panel, elastic (price-responsive) load, inelastic load, and a battery. They individually control elastic load and a battery based on their own system states and retail prices.

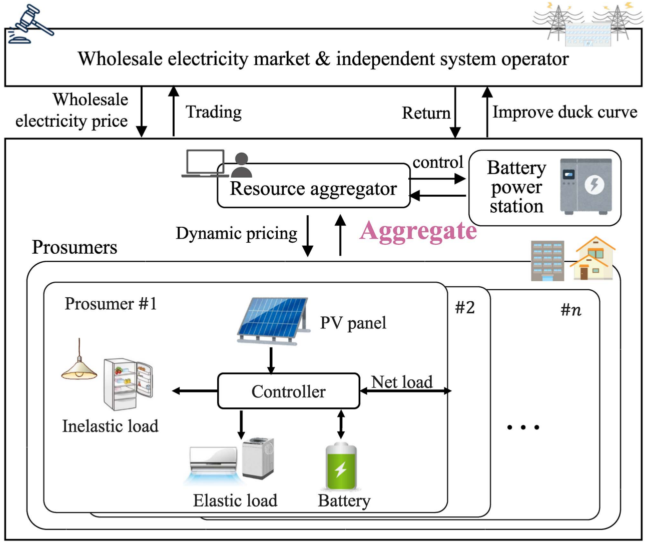


Fig. 2. Overview of hierarchical energy market.

In this study, we focus on the real-time strategy of RA, including dynamic pricing for prosumers and scheduling of a battery power station. We call this problem *Dynamic Pricing and Battery Scheduling (DP-BS)*. As mentioned in Section I, we try to solve the DP-BS problem by employing a model-free DRL approach. We assume that the proposed strategy is implemented in the RA with a DRL-based algorithm and neural networks. In a practical case, the DP-BS by the RA are executed as the following procedures. At each time step, the WEM and the ISO provide the RA with a profile of wholesale electricity prices and requests to flatten the duck curve. Then, the RA will collect information about the net load and the battery SOC from each prosumer. The proposed strategy trained by the DRL algorithm computes the retail prices and the schedule of the RA's battery power station. The battery power station of the RA will operate on the basis of the obtained schedule, and the prosumers decide on schedules of their demand and a battery based on the retail price sent by the RA. After that, the RA aggregates the total net load calculated by summing the power output of the RA's battery power station and the net load of each prosumer and trades it with the WEM and the ISO. Finally, the RA settles the cost of prosumers, the profit of the RA, and the remuneration obtained by flattening the duck curve. All the information that the RA collected is stored in a database of the RA, and the RA trains the neural network based on the DRL algorithm every specific time interval, e.g., 1 week.

The RA's strategy has three main objectives: (1) maximizing revenue from electricity retailing to the prosumers, (2) minimizing the electricity bill of the prosumers, and (3) maximizing the remuneration for the duck curve improvement given by the ISO. Note that the RA's strategy should consider the cost of the prosumer. Neglecting this objective could result in high retail prices, causing discomfort among the prosumers, and potentially leading to the termination of their contract with the RA. On the other hand, the RA can get remuneration by tackling supply-side concerns, such as the duck curve, and it will benefit the RA. Thus, RA has a motivation to improve both supply-side and prosumer benefits, rather than maximizing only RA's profits.

Participation in such programs is valuable to the prosumers. In the dynamic pricing program, electricity retail prices go up and down, depending on the situation, such as the emergence of the duck curve. Prosumers can reduce their electricity costs by scheduling demand and batteries according to changes in retail prices. Since the RA considers the cost minimization of the prosumers in the objectives, it is guaranteed that the retail price will not always be too high. In addition, in our system model, the RA agents receive remuneration from the ISO based on how much net load is flattened. Generally, part of the obtained remuneration will be distributed to the prosumers, and this also helps to reduce the prosumer costs. Finally, this system model to solve the duck curve benefits both the supply and demand sides. In the following section, we describe our model mathematically.

A. Prosumer Model

We assume that the operation of the elastic load and the battery is expected to be automatically or manually controlled in response to the surrounding state and the retail price announced by the RA.

1) Load Model: The load model includes inelastic load and elastic load. The inelastic load refers to base power demand that does not change with respect to retail prices, such as lights, refrigerators, and elevators. The elastic load is a price-responsive load, which is mainly a shiftable load, such as HVAC, clothes washers, tumble dryers, and dishwashers: scheduled operation can be deferred, and the whole elastic load deferred is shifted to a later time slot [41].

While the work in [40] considered the load curtailment, we specifically focus on load deferral, i.e., load shifting to later time slots, to make the prosumer behavior more realistic. In load deferral, the reduced load due to high retail electricity prices, e.g., stopping the operation of washing machines, does not disappear and will be reactivated in a low-price period or after a certain period [42]. Although it is also possible to shift loads to earlier time slots, real-time pricing schemes can pose challenges for the earlier shifting. The time interval between retail price announcements and prosumer demand usage is usually very short, and it is expected that most of the load shifting will be a shift to later time slots. Additionally, it may also be difficult to immediately start up elastic loads that are intended to be used in the future but have not yet been set up. Therefore, this paper specifically only focuses on load deferral (load shifting to later time slots) and explicitly models it as a common load-shifting scenario.

The total load demand $e^{dm}_{t,n}$ of the prosumer n at time t is defined by the inelastic load $e^{inelas}_{t,n}$, the elastic load $e^{elas}_{t,n}$, the deferred load $e^{defer}_{t,n}$ and the shifted load $e^{shft}_{t,n}$, as given by

$$e^{dm}_{t,n} = e^{inelas}_{t,n} + e^{elas}_{t,n} - e^{defer}_{t,n} + e^{shft}_{t,n}. \tag{1}$$

We model the amount of deferred load using price elasticity originating from economic theory [43]. The price elasticity shows the prosumer's sensitivity to the price, which means the percentage change in electricity demand when the price increases by 1%. The deferred load is calculated based on the price elasticity $\xi_{t,n}$ and current price information as follows:

$$e^{defer}_{t,n} = e^{elas}_{t,n} \cdot \xi_{t,n} \cdot \frac{-(\lambda_{t,n} - \mu_t)}{\mu_t}. \tag{2}$$

The price elasticity $\xi_{t,n}$ is usually negative, and a high absolute value means that the end user reacts strongly to the price [44].

The deferred energy $e^{defer}_{t,n}$ is shifted later with a certain probability depending on the current retail price and an elapsed period. To model the shifted load, we introduce an auxiliary variable $u_{t,k,n} \in \{0, 1\}$: when $u_{t,k,n}$ is 1, the deferred energy $e^{defer}_{k,n}$ of prosumer n at time k is scheduled again at time t. Then, the shifted load $e^{shft}_{t,n}$ is modeled as follows:

$$e^{shft}_{t,n} = \sum_{k=1}^{t-1} u_{t,k,n} \cdot e^{defer}_{k,n}. \tag{3}$$

Thus, $e^{shft}_{t,n}$ means the sum of the scheduled energy that is deferred at time k $(1 \leq k \leq t-1)$. The load deferral is characterized by a deadline time at which the shifting must be completed [41]. We model this characteristic by introducing a patience period S_n for each prosumer n [42]. Here, the auxiliary variable $u_{t,k,n}$ is set by the following probability:

$$\mathrm{P}(u_{t,k,n} = 1) = \frac{\lambda^{ub} - \lambda_{t,n}}{\lambda^{ub} - \lambda^{lb}} + \frac{t-k}{S_n}, \tag{4}$$

where $\mathrm{P}(u_{t,k,n} = 1) \in (0, 1)$ is the probability that $u_{t,k,n}$ becomes 1, i.e., the probability that the deferred energy at time k, $e^{defer}_{k,n}$, is shifted to time t. The lower the retail price, the greater the probability that the deferred load will be shifted to the current time slot. At the same time, the longer time period that elapses from the load deferral, the higher the probability due to the effect of patience time. Note that the probability $\mathrm{P}(u_{t,k,n} = 1)$ is clipped so that it ranges from 0 (0%) to 1 (100%).

2) Dissatisfaction Model: The dissatisfaction level of each prosumer is different according to their preferences, and it is generally modeled by a utility function [45], an important concept in microeconomics. There are several types of utility functions; due to its tractability, we use a quadratic function of the amount of deferred load caused by retail prices [46]. The dissatisfaction function $U_{t,n}$ of prosumer n at time t is defined by

$$U_{t,n} = \alpha_n \cdot \left(e^{defer}_{t,n}\right)^2 + \beta_n \cdot e^{defer}_{t,n}. \tag{5}$$

In this equation, the greater the amount of energy deferred, the greater the prosumer's dissatisfaction.

3) Battery Model: The charge/discharge dynamics of the battery of each prosumer are defined by

$$SOC^{bat}_{t+1,n} = \begin{cases} SOC^{bat}_{t,n} + \dfrac{\eta^{bat}_{ch,n} \cdot e^{bat}_{t,n}}{C^{bat}_{n}}, & \text{if } e^{bat}_{t} \geq 0 \\ SOC^{bat}_{t,n} + \dfrac{e^{bat}_{t,n}}{\eta^{bat}_{disch,n} \cdot C^{bat}_{n}}. & \text{otherwise} \end{cases} \tag{6}$$

The charge/discharge energy $e^{bat}_{t,n}$ takes a positive value when charging and a negative value when discharging. The range of charge/discharge energy is constrained by the following constraint.

$$C^{rate,bat}_{disch,n} \cdot C^{bat}_{n} \cdot \Delta t \leq e^{bat}_{t} \leq C^{rate,bat}_{ch,n} \cdot C^{bat}_{n} \cdot \Delta t$$
$$: e^{lb,bat}_{n}, e^{ub,bat}_{n}. \tag{7}$$

Here, $e^{lb,bat}_{n}$ and $e^{ub,bat}_{n}$ mean the upper/lower bounds of the charge/discharge energy defined in (7).

The battery controller of the prosumers also tends to be price-responsive in the typical management problem [47]. In this study, we assume that the prosumers control the battery operation based on a rule-based controller that responds to the retail price. To judge the retail price is high or not, each prosumer sets a price threshold λ^{th}_{n} calculated by

$$\lambda^{th}_{n} = \lambda^{lb} + \left(\lambda^{ub} - \lambda^{lb}\right) \cdot h_{n}. \tag{8}$$

The rule-based controller has four operation modes depending on the current price and PV generation. If the retail price is lower than the threshold, then the system switches to the charge mode, and the prosumer charges the battery with electricity purchased from the RA (mode 1). Otherwise, it switches to discharge mode, and the prosumer discharges the battery to meet the load demand (mode 2). Other rules include stopping the charging/discharging when the battery capacity is full/empty (mode 3) and charging the surplus PV generation to the battery as much as possible regardless of the electricity price (mode 4). To sum up, the charge/discharge energy for each mode is given as follows:

$$e^{bat}_{t,n} = \begin{cases} e^{ub,bat}_{n} \cdot \dfrac{\lambda^{th}_{n} - \lambda_{t,n}}{\lambda^{th}_{t,n} - \lambda^{lb}}, & \text{if mode 1} \\ e^{lb,bat}_{n} \cdot \dfrac{\lambda_{t,n} - \lambda^{th}_{n}}{\lambda^{ub} - \lambda^{th}_{n},} & \text{if mode 2} \\ 0, & \text{if mode 3} \\ e^{pv}_{t,n} - e^{dm}_{t,n}. & \text{if mode 4} \end{cases} \tag{9}$$

Note that the previous study [40] assumed that the prosumer would always charge or discharge the battery at maximum rates whenever retail prices were either below or above a predetermined threshold λ^{th}_{n}. However, this controller did not allow for continuous control of the battery and did not consider current demand and PV generation. In reality, prosumers tend to regulate their battery charge/discharge based on retail price levels and the energy balance between demand and PV generation [47], [48]. To better reflect this behavior, we introduce price sensitivity into the battery control algorithm, as described in equations (8) and (9), so that the amount of energy charged/discharged from the battery is continuously controlled in proportion to the level of retail prices. Furthermore, mode 4 has been included in the equation (9) to account for the prosumer's energy balance.

This battery controller is still simple, but this model is useful as a proof of concept for our proposed DRL method under practical assumptions. The DRL algorithm can be applied and is also considered effective for practical battery controllers due to the model-free nature of DRL, which can learn the prosumer behavior based on observations without specific models.

4) Objective: The objective of the prosumer is to minimize their electricity bill and their dissatisfaction. First, the net load $e^{net}_{t,n}$ of prosumer n at time t is calculated by

$$e^{net}_{t,n} = e^{dm}_{t,n} - e^{pv}_{t,n} + e^{bat}_{t,n}. \tag{10}$$

Here, we denote the positive net load by $e^{net,+}_{t,n}$ and the negative net load by $e^{net,-}_{t,n}$ to distinguish between buying and selling. Finally, the objective of each prosumer is to minimize the total cost $C^{pro}_{t,n}$, defined by

$$\min \sum_{t=1}^{T} C^{pro}_{t,n}, \tag{11}$$

$$C^{pro}_{t,n} = \lambda_{t,n} \cdot e^{net,+}_{t,n} - \kappa_{t} \cdot e^{net,-}_{t,n} + U_{t,n}. \tag{12}$$

The first term means the electricity cost, the second term is the revenue from selling the surplus energy, and the third term is the dissatisfaction of the prosumer.

B. Resource Aggregator Model

1) Pricing Model: We assume that the RA sells electricity to each prosumer at a retail price that varies over time $\lambda_{t,n}$ and buys surplus electricity from the prosumers at a purchase price κ_{t}. The retail prices $\lambda_{t,n}$ also vary for each prosumer, and the purchase price κ_{t} is generally equal to the wholesale electricity price μ_{t}. Then the RA trades the aggregated electricity at a wholesale electricity price μ_{t} notified at every time step by the WEM. We introduce a constraint for retail prices to avoid unfair pricing to the prosumers, given by

$$\nu \cdot \mu_{min} \leq \lambda_{t,n} \leq \nu \cdot \mu_{max} : \lambda^{lb}, \lambda^{ub}. \tag{13}$$

The retailed electricity price $\lambda_{t,n}$ for prosumer n at time t is decided by our proposed strategy.

2) Battery Power Station Model: The RA controls the charge/discharge amount of the battery power station to provide flexibility to the total net load. The dynamics of the battery power station are specified by the following equation as well as the prosumer's battery:

$$SOC^{ra}_{t+1} = \begin{cases} SOC^{ra}_{t} + \dfrac{\eta^{ra}_{ch} \cdot E^{ra}_{t}}{C^{ra}}, & \text{if } E^{ra}_{t} \geq 0 \\ SOC^{ra}_{t} + \dfrac{E^{ra}_{t}}{\eta^{ra}_{disch} \cdot C^{ra}}, & \text{otherwise} \end{cases} \tag{14}$$

$$C^{rate,ra}_{disch} \cdot C^{ra} \cdot \Delta t \leq E^{ra}_{t} \leq C^{rate,ra}_{ch} \cdot C^{ra} \cdot \Delta t. \tag{15}$$

The charge/discharge energy E^{ra}_{t} takes a positive value when charging and a negative value when discharging. The charge/discharge energy E^{ra}_{t} is also controlled by our proposed strategy.

3) Objective: The objective of the RA is to maximize the RA's profit, minimize the prosumer's cost, and improve the duck curve. The RA gets profits from the electricity trade, and the total net load E_t^{net} is defined by

$$E_t^{net} = \sum_{n=1}^{N} e_{t,n}^{net} + E_t^{ra}. \tag{16}$$

Finally, the objective function of the RA is defined by

$$\max \sum_{t=1}^{T} \omega_1 \cdot P_t^{ra} - \omega_2 \cdot \sum_{t=1}^{T}\sum_{n=1}^{N} C_{t,n}^{pro} + (1-\omega_1-\omega_2) \cdot \sum_{t=1}^{T} R_t^{duck}, \tag{17}$$

$$P_t^{ra} = \sum_{n=1}^{N}\left(\lambda_{t,n} \cdot e_{t,n}^{net,+} - \kappa_t \cdot e_{t,n}^{net,-}\right) - \mu_t \cdot E_t^{net}. \tag{18}$$

The weight parameters ω_1 and ω_2 are used to adjust the importance of each objective term, and these range from 0 to 1. To improve the duck curve, the design of R_t^{duck} is very important. We give the detailed formulation of R_t^{duck} in Section III.

C. Dynamic Pricing and Battery Scheduling Problem

We can formulate the central (cooperative) DP-BS problem in the RA as follows:

$$\begin{aligned} \max_{\mathbf{v}} \quad & (17) \\ \text{s.t.} \quad & (13) - (16), (18), \quad \forall t, n \\ & \mathbf{w_{t,n}} = \arg\min_{\mathbf{w_{t,n}}} \sum_{t=1}^{T} C_{t,n}^{pro}, \quad \forall t, n \\ \text{s.t.} \quad & (1) - (9), (10), (12), \quad \forall t, n. \end{aligned} \tag{19}$$

This is a bi-level problem: the upper level optimizes the decision for the retail prices and battery power station, and $\mathbf{v}$ is a set $\{\forall t, n : \lambda_{t,n}, E_t^{ra}\}$. The lower level optimizes the behavior of the prosumer, and the decision set $\mathbf{w_{t,n}}$ is $\{e_{t,n}^{bat}, e_{t,n}^{shifted}\}$. A conventional solution for such a bi-level problem is a mathematical program with equilibrium constraints (MPEC) approach [49]. In this approach, a bi-level problem is reformulated as a single-level optimization problem by converting the lower-layer problem into its equilibrium conditions, generally referred to as the Karush-Kuhn-Tucker (KKT) conditions. The KKT conditions of the lower-layer problem, which consists of additional variables, constraints, and complementarity conditions, are then incorporated into the upper-layer problem. The lower-layer problem is removed from the bi-level problem, and the resulting single-level optimization problem can be solved by using a standard nonlinear programming (NLP) solver.

However, solving the problem (19) with an MPEC approach is challenging for the following reasons. First, solving this optimization problem requires uncertain future information such as wholesale electricity prices and net load. This information is essentially unknown in advance, and associated forecast errors result in high costs. Second, the RA does not have access to detailed models of prosumers due to privacy concerns. Without the prosumer response model, the solution obtained by solving the optimization problem (19) will be unreliable. Third, solving a bi-level optimization problem using MPEC can be computationally expensive because it involves solving a highly nonlinear optimization problem with additional variables and constraints. Thus, solving the bi-level problem (19) using an MPEC approach is not realistic in practice. We address these issues by applying a model-free DRL approach.

III. Deep Reinforcement Learning-Based Strategy

A. Overview

To solve the DP-BS problem, we employ a model-free DRL algorithm. The reasons for using DRL to solve the DP-BS problem are twofold: its adaptive capability and model-free nature. First, in a DRL paradigm, learning policy proceeds adaptively in response to changes in the dynamic environment, taking into account uncertainties, e.g., wholesale prices and net load change. Second, DRL methods can learn an optimal policy to make a decision through observable interactions without detailed system models. This model-free nature requires no knowledge of the detailed system models of each prosumer, i.e., privacy concerns can be resolved. Moreover, once trained, decision-making by DRL takes negligible computation time, typically less than one second, without the need to solve complex problems like NLP.

In a DRL problem, a decision maker is called an agent, while a surrounding follower interacting with the agent is called an environment. The agent-environment interactions must be modeled as a Markov Decision Process (MDP) to apply DRL methods. The MDP consists of a set of a state, an action, a transition probability, and a reward function [50]. Following the probability of transition, the state of the environment s moves to a new state s' under an action of the agent a. The reward is a numerical score that evaluates whether the action taken is good or not. To choose appropriate actions, the agent learns a policy π_θ parameterized by θ, which is a way of decision-making that maximizes the reward function. The typical procedure of the RL framework at time t is for the agent to take action a_t in the environment based on the policy and then feed back the reward r_t and the new state s_{t+1} to the agent. The DRL algorithm updates the policy based on the transition information (s_t, a_t, r_t, s_{t+1}).

B. Formulation of Markov Decision Process

We reformulate the DP-BS problem as the MDP to handle the problem using the DRL algorithm. Fig. 3 shows our DRL framework for the DP-BS problem. The agent is the RA, and the environment is the ISO with the WEM, the battery power station, and the prosumers. Unlike the central NLP problem (19), our framework only optimizes the retail prices and the usage of the RA's battery power station. Note that the formulation does not require a model of the transition probabilities since the proposed method is a model-free method that learns from data.

1) State: The state observations consist of seven types of information for the ISO, the prosumers, and the battery power

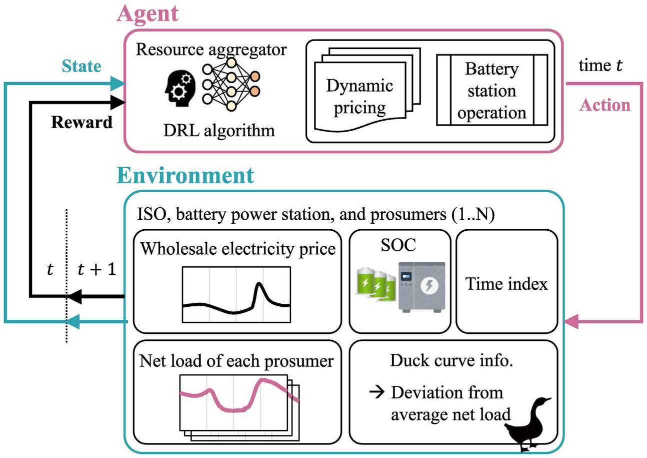

Agent
State
Reward
Resource aggregator
DRL algorithm
Dynamic pricing
Battery station operation
time t
Action
Environment
ISO, battery power station, and prosumers (1..N)
Wholesale electricity price
SOC
Time index
t
$t+1$
Net load of each prosumer
Duck curve info.
→ Deviation from average net load

Fig. 3. Illustration of DRL framework for DP-BS problem.

station: time slot index $t' = t \bmod T$, the wholesale electricity price μ_t, the net load for each prosumer $e^{net}_{t,n}$, the sum of the remaining deferred energy $\Sigma^{t-1}_{k=1} e^{defer}_{k,n}$, the SOC of prosumer's battery / battery power station $SOC^{bat}_{t,n}$ / SOC^{ra}_t, and the deviation E^{dev}_t between the total net load and daily average net load calculated by the following equations.

$$E^{dev}_t = \sum_{n=1}^{N} e^{net}_{t,n} - E^{avg}_{day}, \tag{20}$$

$$E^{avg}_{day} = \sum_{t=1}^{T} \sum_{n=1}^{N} \frac{e^{net}_{t,n}}{N}. \tag{21}$$

We assume the value of the daily average of the net load E^{avg}_{day} is known in advance. Although the specific forecasting method for E^{avg}_{day} is out of the scope of this paper, we can directly forecast E^{avg}_{day} using short-term forecasting that has already been developed in the work [51], which has high accuracy.

Compared to the previous work [40], the addition of two important state variables, $\Sigma^{t-1}_{k=1} e^{defer}_{k,n}$ and E^{dev}_t, has been made to along with the introduction of load shifting capability and support the training of the RA agent. The variable $\Sigma^{t-1}_{k=1} e^{defer}_{k,n}$ helps the RA agent to understand how many demand loads are already deferred and will be shifted, while E^{dev}_t provides insight into the state of the duck curve and total net load. Accordingly, we define the state s_t at time t by

$$s_t = \Big(t', \mu_t, e^{net}_{t,1}, \ldots, e^{net}_{t,N}, \Sigma^{t-1}_{k=1} e^{defer}_{k,n}, \ldots, \Sigma^{t-1}_{k=1} e^{defer}_{k,n}, SOC^{bat}_{t,1}, \ldots, SOC^{bat}_{t,N}, SOC^{ra}_t, E^{dev}_t\Big). \tag{22}$$

2) Action: The RA decides the retail price for each prosumer $\lambda_{t,n}$ and the operation of the battery power station E^{ra}_t, aiming to maximize the objectives. We assume that the action space of both variables is continuous and that its range is constrained by the upper/lower bound as given in (13) and (15). Thus, the actions of RA a_t at time t are given by

$$a_t = \big(\lambda_{t,1}, \ldots, \lambda_{t,N},\ E^{ra}_t\big). \tag{23}$$

TABLE I
PROPOSED REWARD TERMS FOR IMPROVING DUCK CURVE

Reward	Description
$R^{avg}_t = (E^{net}_t - E^{avg}_{day})^2$	Quadratic penalty of deviation between current net load and daily average net load
$R^{diff}_t = (E^{net}_t - E^{net}_{t-1})^2$	Quadratic penalty of net load difference for consecutive time slots
$R^{quad}_t = (E^{net}_t)^2$	Quadratic penalty of total net load
$R^{no}_t = 0$	No reward for duck curve improvement

3) Reward Function: The objective of the RA agent is to maximize the RA's profit, minimize the prosumer's cost, and maximize the duck curve improvement. Hence, we define the following reward r_t at time t similar to the problem (19):

$$r_t = \omega_1 \cdot P^{ra}_t - \omega_2 \cdot \sum_{n=1}^{N} C^{pro}_{t,n} - (1 - \omega_1 - \omega_2) \cdot R^{duck}_t. \tag{24}$$

Note that the value of the weights ω_1 and ω_2 should be carefully chosen based on each entity's preferences. The adjustment method of the weights is outside the scope of this paper; however, we verify the effect of the weight value choices in Section IV-E.

The design of an appropriate reward function is critical to training and deploying the DRL agent efficiently. Here, we show four different reward terms as R^{duck}_t to improve the duck curve (Table I). The content of the duck curve is a large peak valley deviation of the net load and a steep change of net load for consecutive time slots. The power generation cost of power plants for flexibility is typically defined as a quadratic function of the net load [52]. Thus, we propose reward terms as quadratic penalty functions for deviation from a daily average, the net load difference for consecutive time slots, and the total net load. In addition, to compare performance, we also give the case with no reward term for the improvement of the duck curve R^{no}_t. We verify the proposed terms in Section IV.

C. Algorithm Design

We train the agents to solve the DP-BS problem using Proximal Policy Optimization (PPO) [53], which is one of the state-of-the-art DRL algorithms. This is because the performance of the PPO algorithm is compatible with or better than other state-of-the-art DRL algorithms in the DRL benchmark for tasks with a continuous action space [53].

PPO is an actor-critic policy gradient method parameterized by neural networks and improves the stabilization of learning by preventing a large policy update. To do this, the ratio of the old to the new policy is clipped, and a lower update bound, i.e., a pessimistic bound, is chosen. At the k-th iteration, the parameter θ of a policy π_θ is updated by

$$\underset{\theta}{\text{maximize}}\, \hat{\mathbb{E}}\Big[L^C(\theta)\Big], \tag{25}$$

$$L^C(\theta) = \min\left(\begin{array}{l} \frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}(s,a), \\ \text{clip}\Big(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}, 1-\epsilon, 1+\epsilon\Big) \cdot A^{\pi_{\theta_k}}(s,a) \end{array}\right), \tag{26}$$

Algorithm 1: Training and Deployment Process of Proposed DP-BS Strategy Based on PPO Algorithm

```
Initialize policy parameters $\theta$;
Initialize training memory $\mathcal{M}$;
/* Training process                                        */
for $episode = 1, max\_episode$ do
    Initialize the state of the hierarchical energy market; for $t$ = 1, $T$ do
        Receive initial state $s_t$;
        Sample action $a_t$ according to $\pi_{\theta_k}$;
        Calculate reward $r_t$ and observe new state $s_{t+1}$;
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in training memory $\mathcal{M}$;
    Get a mini-batch from training memory $\mathcal{M}$;
    Estimate advantages $A^{\pi_{\theta_k}}$ using any advantage estimation algorithm;
    Optimize loss function $L^{PPO}$ given by (29);
    Update $\theta$ with any gradient optimizer;
/* Deployment process                                      */
Deploy the trained policy $\pi_\theta$ for the RA;
Perform the DP-BS in the actual system based on $\pi_\theta$;
```

where Ê denotes the empirical expectation over time steps and L^C is a surrogate objective. Here $A^{\pi_{\theta_k}}$ is the estimated advantage in the k-th iteration and ϵ is a hyperparameter that denotes the clipping range. The function clip() clips the policy update ratio within $[1-\epsilon, 1+\epsilon]$.

PPO is generally implemented with a neural network architecture that shares parameters between the policy and value functions. Here, the value function $V_{\theta_k}(s)$ parameterized by θ_k in the critic network is updated with respect to the mean square error of the value function L^{VF}:

$$\underset{\theta}{\text{minimize}}\, \hat{\mathbb{E}}\big[L^{VF}(\theta)\big], \tag{27}$$

$$L^{VF}(\theta) = \big(V_\theta(s) - V^{targ}(s)\big)^2, \tag{28}$$

where $V^{targ}s$ is the target value of an old value function. Finally, the loss function of PPO L^{PPO} to maximize is the sum of L^C, L^{VF}, and an entropy bonus S:

$$L^{PPO}(\theta) = \hat{\mathbb{E}}\Big[L^C(\theta) - c_1 \cdot L^{VF}(\theta) + c_2 \cdot S[\pi_\theta(s)], \tag{29}$$

where c_1 and c_2 are coefficients.

The pseudocode of the proposed strategy is presented in Algorithm 1. The RA agents in the PPO algorithm are first trained repeatedly, and then the trained agents are deployed in the RA and operate the DP-BS in real-time.

IV. Simulation Results

In this section, we present several simulation experiments to evaluate the proposed DP-BS strategy. The experimental setup and implementation details are first described, and then we compare the performance of the proposed strategy with other baselines in terms of the improvement of the duck curve and computational complexity. In addition, different scenarios are performed to access the system performance of the proposed strategy.

A. Experimental Setup

We assume that the hierarchical energy market consists of an ISO, an RA, and ten prosumers ($N = 10$). The interval of dynamic pricing and battery scheduling by the RA was set to 30 min, and an episode length was set to a day, i.e., $T = 48$. The wholesale electricity prices were obtained for the entire year of 2017 from a California ISO [54]. The RA parameters are given in Table II. The purchase price κ_t is assumed to be the same value as the wholesale electricity price μ_t at that time. Both the weights of ω_1 and ω_2 in (29) were set to 0.2 so that the duck curve improvement would be considered more important than other terms. However, the effect of weight is also explored in Section IV-E.

TABLE II
Parameter Setting for RA and Prosumers

Parameter	Symbol	Value
Coefficient for price limit	ν	1.5
Battery capacity [kWh]	C^{ra}	300
Charging efficiency	η^{ra}_{ch}	0.9
Discharging efficiency	η^{ra}_{disch}	0.9
Maximum charging power rate	$C^{ra,rate}_{ch}$	0.3
Maximum discharging power rate	$C^{ra,rate}_{disch}$	0.3
Patience period [30 min]	S_n	6, 12, or 18
Coefficient for utility function	α_n	Rand. value within 1∼4
	β_n	1
Battery capacity [kWh]	C^{bat}_n	10, 15, or 20
PV panel size [kW]	-	10, 15, or 20
Charging efficiency	$\eta^{bat}_{ch,n}$	0.9
Discharging efficiency	$\eta^{bat}_{disch,n}$	0.9
Maximum charging power rate	$C^{bat,rate}_{ch,n}$	0.3
Maximum discharging power rate	$C^{bat,rate}_{disch,n}$	0.3
Coefficient for price threshold	h_n	0.5
Average total net load [kW]	-	350
Average ratio of inelastic load to total demand [%]	-	60
Average ratio of elastic load to total demand [%]	-	40

On the other hand, the prosumers are simulated using the building's energy consumption profiles collected by the Building Data Genome Project 2 [55] and the PV generation profiles provided by the California Distributed Generation Statistics [56]. The period under study for both profiles covers the entire year of 2017. Note that we resampled these datasets at 30-minute intervals and normalized them according to the building site area. The other parameters of the prosumers are also given in Table II. The price elasticity profile $\xi_{t,n}$ ranges from –0.2 to –0.8 based on the literature [44], and we manually generated it.

The evaluation metrics to improve the duck curve are the average of the standard deviation of the net load, denoted by std, and the peak-to-average ratio (PAR) of the total net load E^{net}_t for each day, as given by:

$$std = \sqrt{\frac{\sum_{t=1}^{T} |E^{net}_t - \sum_{t=1}^{T} E^{net}_t / T|^2}{T-1}}, \tag{30}$$

$$PAR = \frac{\max\big(E^{net}_t\big)}{\sum_{t=1}^{T} E^{net}_t / T}, \tag{31}$$

where $\max(E^{net}_t)$ is a function that finds a maximum value for the total net load.

B. Implementation and Training Process

We implemented a simulator and a PPO algorithm in `python`. The simulator was built on the OpenAI Gym

TABLE III
PARAMETERS FOR PPO ALGORITHM

Parameter	Value
Number of environments in parallel	16
Number of episodes max_episode	20,000
Batch size	128
Training memory size $\mathcal{M}$	2048
Number of epochs	5
Clip range ϵ	0.1
Discount factor	0.995
Learning rate	0.002
Value function coefficient for loss function c_1	0.5
Entropy bonus coefficient for loss function c_2	3.6e-8
Number of hidden layers	2
Number of neurons	[256,256]
Activation function	ReLu

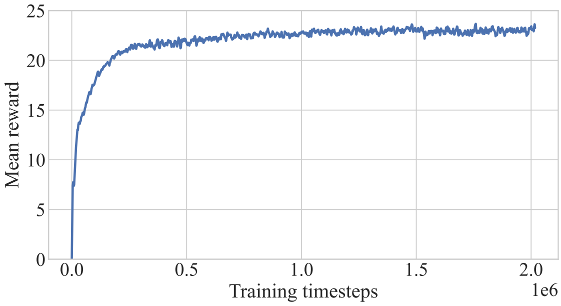
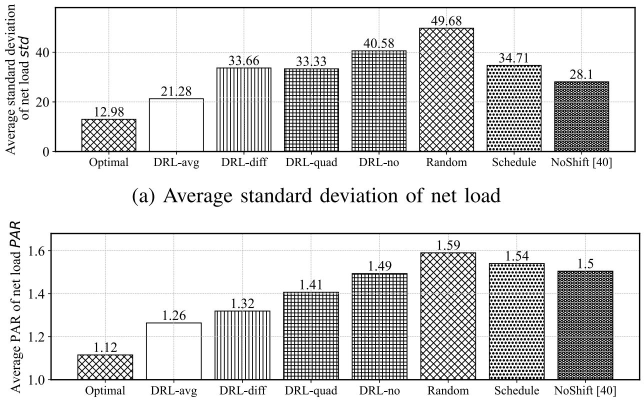


Fig. 4. Typical training curve of proposed method.

framework [57], which allows us to easily observe the agent-environment interaction. Furthermore, the proposed DP-BS strategy was implemented using the python library `stable-baselines3` [58], which is an open-source DRL framework. The parameters of the PPO were fine-tuned using `Optuna` [59], and the obtained parameters are shown in Table III. In our setting, the neural network is shared for both policy and value functions in actor-critic.

It is worth mentioning that we made some implementation techniques for the DRL framework to stabilize the training and shorten the training time. We normalized the value of state observation and actions within $[-1, +1]$ using min-max normalization. This state-action normalization is required to ensure that the neural network is not too dependent on the scale of the features [60]. The scale of the reward terms in the equation (24) is aligned based on the standard score, which is calculated by the difference between the value and the mean divided by the standard deviation. This standardization of rewards ensures that the RL agents can be tuned with similar hyperparameters. Firstly, hyperparameter tuning is carefully performed using the previous year's data for representative parameters that include the number of neurons, the number of parallel environments, learning rate, and discounted factor as important hyperparameters [61], and the same parameters are applied for all experiments with the aforementioned normalization and standardization technique.

The configurations of the test-bed machine include Intel Core i7-10700 CPU @ 2.90 GHz, Nvidia GeForce RTX 2080 Ti GPU, and 16 GB DDR4 RAM. The average execution time of training the neural network is 29.4 min for 2M steps, and the typical training curve is shown in Fig. 4. After repeated training with the simulator using a one-month dataset, the trained agent is deployed and makes a decision for the DP-BS in real time.



(a) Average standard deviation of net load

(b) Average PAR of net load

Fig. 5. Results of average standard deviation and average PAR of total net load for one week's simulation from August 1 to August 7 compared to baselines and DRL-based method with different reward functions.

C. Comparison With Baseline Methods

First, the proposed strategy is compared with representative baseline methods to evaluate the performance of the duck curve improvement. In all, we compare seven methods:

- *Optimal:* Conventional MPEC approach [49] assuming ideal scenarios; the bi-level problem (19) that optimizes R_t^{avg} is transformed into a single-level NLP problem using an MPEC approach. The resulting NLP problem covers a 24-hour planning period with a 30-minute resolution and is solved daily. This method assumes the ideal scenario of DP-BS, where all future inputs and system configurations are known in advance, and all operations, including the prosumer (the elastic load and the battery) and the RA (the retail prices and the battery power station), are under control. The solution obtained is ideal and shows the maximum potential of DP-BS.
- *DRL-avg:* Proposed DRL-based strategy using R_t^{avg}.
- *DRL-diff:* Proposed DRL-based strategy using R_t^{diff}.
- *DRL-quad:* Proposed DRL-based strategy using R_t^{quad}.
- *DRL-no:* Proposed DRL-based strategy using R_t^{no}.
- *Random:* Randomly determines the retail price and the battery operation.
- *Schedule:* Pre-defined schedule; retail prices are set to 75% of the price range (13) in peak hours (4 p.m.–9 p.m.), otherwise set to 25% of that. The battery station discharges at a constant rate of 0.2 C during the same peak hours and charges at 0.1 C during another period.
- *NoShift [40]:* DRL-based strategy proposed in the previous work [40]; the prosumer has no load shifting (load deferral) mechanism (only reducing demand according to retail prices) and always charge/discharge the battery at maximum rates as described in Sections II-A1 and II-A3, respectively. The MDP model do not include important states, $\Sigma_{k=1}^{t-1} e_{k,n}^{defer}$ and E_t^{dev}, as explained in Section III-B. The reward function is the same as DRL-avg.

Figs. 5(a) and 5(b) depict the standard deviation and PAR of the total net load for a seven-day span from August 1st to

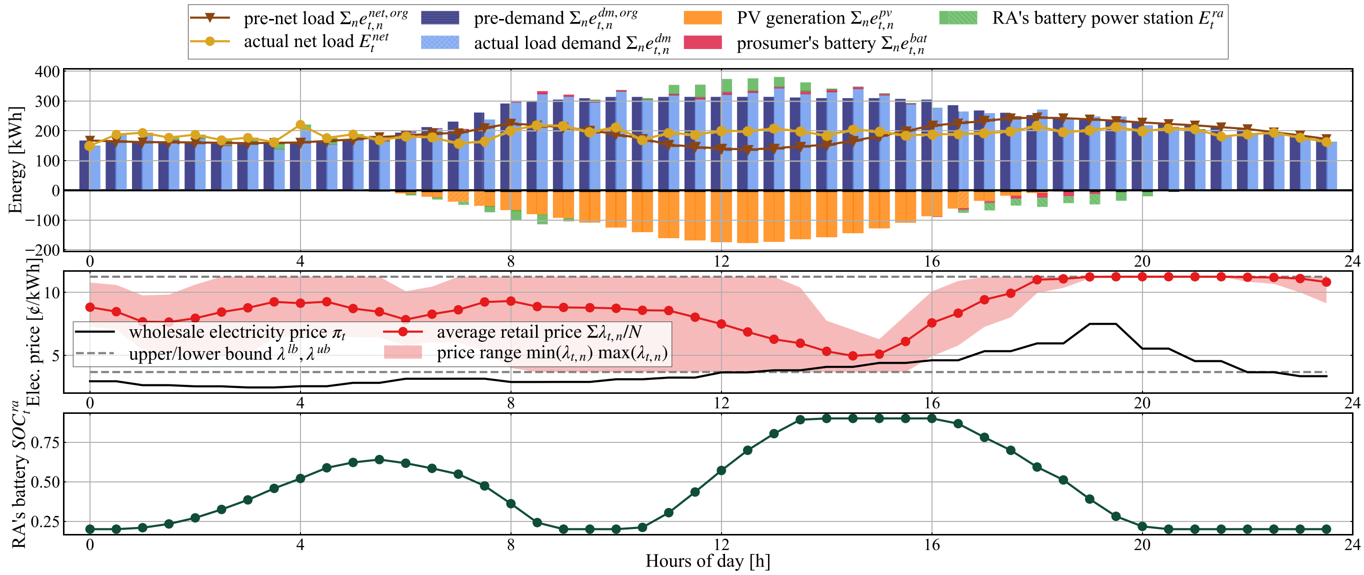


Fig. 6. Result profile for August 5 using DRL-avg; (upper) energy and net load change every 30 min, (middle) wholesale electricity prices and average retail price for all prosumers with shaded area, indicating max-min values of retail prices at each time step, and (lower) the SOC of the RA's battery.

August 7th, 2017. The DRL methods were trained using the simulation and input profiles from the preceding month of July 1st to July 31st. Both figures reveal that, with the exception of the Optimal method, DRL-avg exhibited the superior standard deviation and PAR. In Fig. 5(a), DRL-avg improved the standard deviation of the total net load by a minimum of 24.2% compared to the NoShift method, and up to 57.1% in comparison to the Random method. Similarly, Fig. 5(b) illustrates that DRL-avg enhanced the PAR of the total net load by at least 6% when compared to DRL-diff and up to 23% relative to Random. The reason behind the proposed strategy's superiority over the NoShift method lies in its ability to improve the duck curve through the introduction of load shifting, or load deferral. Without load shifting, the period of low net load, i.e., the valley of the duck curve, remains unfilled, leading to increased standard deviation when utilizing the NoShift method. In addition, the NoShift's battery controller caused the prosumer's battery SOC to drain quickly by always charging and discharging at maximum rates, thus their battery did not contribute to the improvement in the duck curve. Furthermore, the NoShift method lacks knowledge of the deviation between the total net load and the daily average net load, E^{dev}_t, resulting in increased maximum net load and PAR. Note that Optimal is an ideal baseline that has complete future knowledge and system models. Even though the proposed DRL-based strategy observes only the current state of the system, the DRL-avg improved the duck curve in terms of both metrics, and its values were closest to Optimal.

Fig. 6 shows the detailed profiles on August 5 using DRL-avg. The upper figure shows the aggregated energy profiles of the RA and the prosumers, where the pre-net load, which is the net load scheduled originally, is calculated by $e^{net,org}_{t,n} = e^{inelas}_{t,n} + e^{elas}_{t,n} - e^{pv}_{t,n}$. The pre-demand, which is the demand scheduled originally, is calculated by $e^{dm,org}_{t,n} = e^{inelas}_{t,n} + e^{elas}_{t,n}$. In the upper figure, there are two bars next to each other every 30 min. The left bars are the energy profiles scheduled originally, and the right bars are the energy profiles after the DP-BS. The positive value means the energy demand including the total demand and battery charging, and the negative energy means the energy supply by PV generation and battery discharging. From the upper figure, comparing the pre-net load $\sum_n e^{net,org}_{t,n}$ to the actual net load E^{net}_t, the actual net load becomes larger than the pre-net load around the noon (11 a.m. – 3 p.m.) and smaller in the morning and the evening (7 a.m. and 4 p.m. – 8 p.m.). This shows that the duck curve that appeared in the pre-net load is smoothed out. The improvement of the duck curve resulted from the following two factors. First, the prosumer contributed to flattening the net load. The actual load demand $\sum_n e^{dm}_{t,n}$ is larger than the pre-net load $\sum_n e^{dm,org}_{t,n}$ around the noon (11 a.m. – 3 p.m.), comparing the two adjacent blue bars in the figure. This means that the prosumer shifted their demand to noon. At the same time, the prosumer charged the battery in the daytime and discharged it in the evening (4 p.m. – 7 p.m.), as represented by the red bars in the figure. Second, the RA's battery power station, which is shown by the green bars in the figure, was scheduled to charge around noon and discharge in the morning and evening. Moreover, the middle figure represents the results of dynamic pricing, and the average prices for all prosumers are plotted. The shaded area means the price deviation among prosumers, and we can see that the trend of the retail prices is similar for each prosumer. The retail prices were set relatively low from noon to 3 p.m. As a result, the prosumers are encouraged to shift their demand and charge the battery in the period 12 p.m. – 3 p.m. Finally, the lower figure shows the SOC profiles of the RA's battery. DRL-avg method learned to charge when the pre-net load is relatively low (11 a.m. – 2 p.m.) and discharge when the pre-net load is relatively high (4 p.m. – 8 p.m.).

Fig. 7 shows the detailed profiles of prosumer #8 for August 6 to demonstrate how the demand and battery respond to the retail price. Similar to Fig. 6, in the upper figure, the left bars are the demand profiles scheduled originally, and the

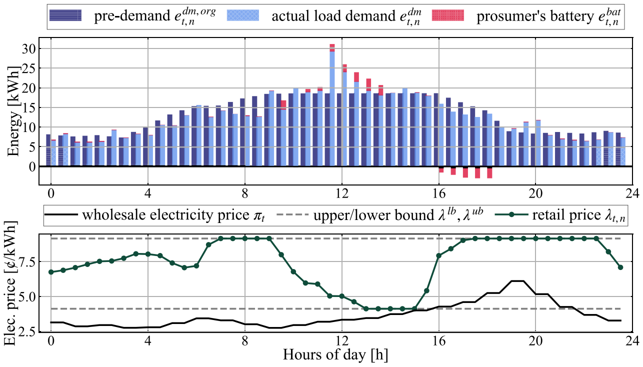


Fig. 7. Detail profile of prosumer #8 for August 6 using DRL-avg; (upper) demand change and battery energy, (lower) retail price and wholesale price.

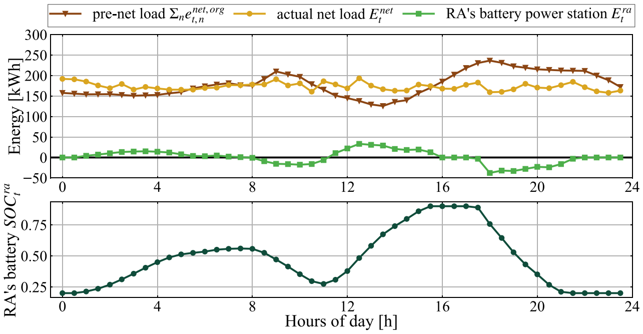


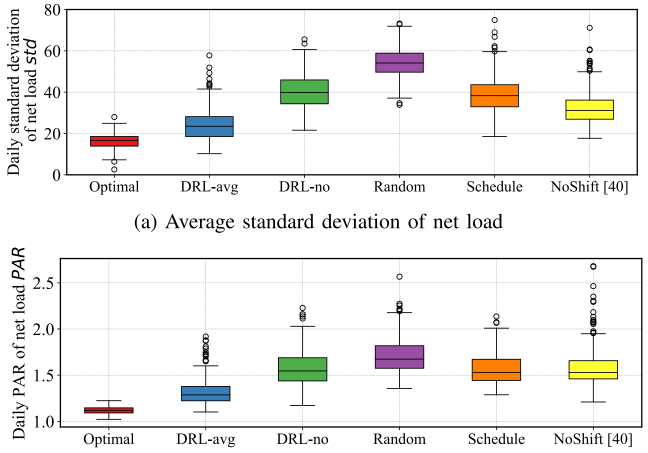
Fig. 8. Result profile of RA for August 2 using DRL-avg; (upper) total net load change and energy of RA's battery power station, (lower) SOC of RA's battery power station.

right bars are the actual demand and battery profiles after the retail price announcement; the lower figure shows the retail price of prosumer #8 given by the RA. Firstly, the RA raised the retail price from midnight to morning, decreased it until 3 p.m., and then raised the retail price again. According to the price change, the prosumer #8 shifted their demand from morning (especially around 8 a.m.) to daytime where PV generation is large. As for the battery, the prosumer charged the battery during the lower retail price period (9 a.m. – 1 p.m.) and discharged it during the higher period (4 p.m. – 6 p.m.). Finally, the RA decided the retail price that made the prosumer increase their net load around noon and reduce it during peak hours for improving the duck curve.

Fig. 8 shows the profiles of the net load and the RA's battery power station on another day (August 2). The results show that the DRL agent operated the RA's battery to discharge during peak period (9 a.m. and 6 p.m. – 9 p.m.) while charging to fill the net load valley until the battery capacity was full (11 a.m. – 3 p.m.). From these results, the proposed strategy with the appropriate reward function can make an effective decision on the retail prices and the RA's battery schedule in real time for improving the duck curve.

D. Whole-Year Simulation

To evaluate the annual performance of our proposed DRL strategy, we conduct a whole-year simulation using five methods: Optimal, DRL-avg, DRL-no, Random, and Schedule. In the DRL methods, we iterated the agent training every week using data from the previous month corresponding to that week. This iteration allows the agent to consider seasonal changes in the target system, such as wholesale electricity prices and net load change.



(a) Average standard deviation of net load

(b) Average PAR of net load

Fig. 9. Results of average standard deviation and average PAR for the 2017 whole-year simulation compared to baselines.

Fig. 9(a) and Fig. 9(b) are box plots showing the daily standard deviation and the daily PAR of the total net load for all of 2017. Consequently, the results show the same trend as in Fig. 5. Without future information and the direct control of the prosumer, the performance of DRL-avg is close to Optimal. Furthermore, DRL-avg also outperforms NoShift [40] for both metrics. DRL-avg achieves a reduction in both the standard deviation and the PAR compared to other baselines. In the best case of DRL-avg, the standard deviation was 10.23, and the PAR was 1.1, which are improvements of 73% and 21% over Random on the same day. We confirm that DRL-avg has the potential to improve the duck curve throughout the year.

E. Effect of Weight Coefficient

This section discusses the effect of a weight coefficient in the reward function (24) on system performance. The simulation period is one week from August 1 to 7 using the DRL-avg method. Both weight coefficients of the RA's profit ω_1 and the prosumer's cost ω_2 changed from 0.0 to 1.0 subject to $\omega_1 + \omega_2 \leq 1$, and the weight of the duck curve improvement was calculated by $(1 - \omega_1 - \omega_2)$. The large weights mean that the corresponding term is considered important.

Fig. 10 shows the heat maps of the system performance with different weights over one week: the standard deviation of net load *std*, PAR of net load, total RA profit, and total prosumer cost. The lower left on the heat map, the greater the importance of duck curve improvement $(1 - \omega_1 - \omega_2)$. The light (yellow) squares mean better values, and the dark (navy) squares mean worse values. As can be seen from the *std* and PAR results, the performance trends of the standard deviation and PAR with respect to the weights are similar. A greater weight of RA's profit increases the standard deviation and PAR of the net load. In terms of the duck curve improvement, we can see that the

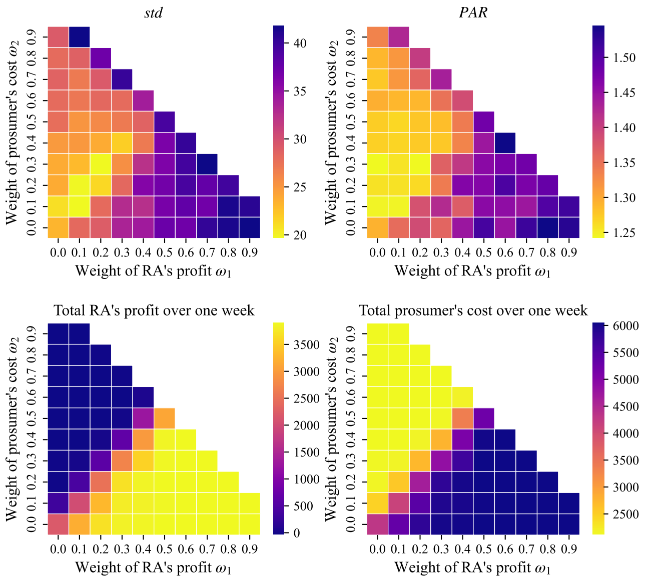
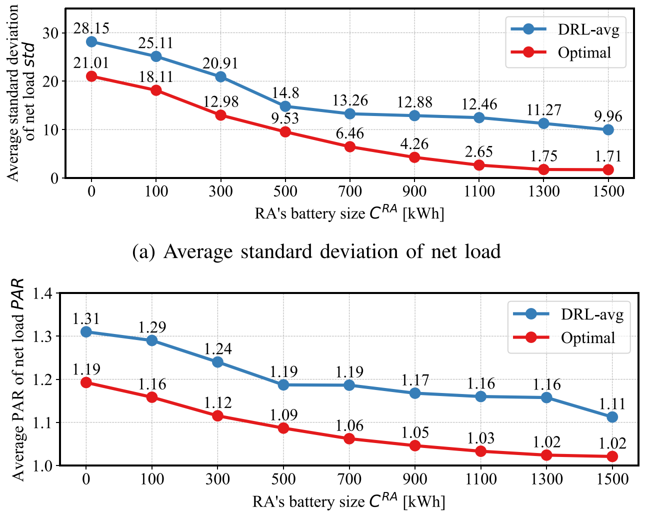


Fig. 10. Results of system performance with different weights ω_1, ω_2 from Aug. 1 to 7 / Weight of duck curve improvement, standard deviation of net load, PAR of net load, total RA's profit, and total prosumer's cost are shown.



(a) Average standard deviation of net load

(b) Average PAR of net load

Fig. 11. Results of average standard deviation and average PAR for one week simulation from Aug. 1 to 7 with different battery sizes for DRL-avg and Optimal.

TABLE IV
COMPUTATIONAL TIME FOR TRAINING AND EXECUTION WITH DIFFERENT NUMBERS OF PROSUMERS

# of prosumers	Optimal (NLP)	Proposed (DRL)	
	Average CPU time at execution (online) [h]	Training time to converge (offline) [h]	Average CPU time at execution (online) [s]
10	0.28	0.16	2.28e-4
20	1.44	0.33	2.31e-4
30	2.94	0.56	2.34e-4
40	4.89	0.77	2.37e-4
50	7.57	0.96	2.39e-4

best weight pair is $(\omega_1, \omega_2, 1-\omega_1-\omega_2) = (0.2, 0.3, 0.5)$. Furthermore, the relationship between the RA's profit and the prosumer's cost is clearly a trade-off that can be controlled by adjusting the weight coefficients. Because both of them tend to be extreme values, the weights of RA's profit and prosumer's cost should be chosen to be as equal as possible. As a result, we can choose the preferred operating point by referring to these heat maps and adjusting the weight coefficient.

F. Effect of Battery Power Station Size

In this section, the effect of the battery power station's size on performance is demonstrated for the duck-curve improvement. The simulation period is one week from August 1 to August 7 using the DRL-avg method. The size of the battery power station increased from 0 to 1500 kWh. The results are compared to the Optimal method, which has complete control and perfect knowledge of future information, as described in Section IV-C. In this simulation, we increased the number of episodes from 2M to 3M steps to ensure convergence as the battery size increases and the action space expands. The average training time is 46.1 min, which is acceptable for our weekly update assumptions.

Fig. 11(a) shows the standard deviation of the total net load for each battery size, while Fig. 11(b) shows the PAR of the total net load for each battery power station size. As the battery power station capacity increases, both the Optimal and DRL-avg methods show a decreasing trend in the standard deviation and PAR. The Optimal solution reaches its peak between 1300-1500 kWh, while the DRL-avg methodology exhibits a similar trend but with a performance gap between the two methods. This is because the Optimal method assumes complete control over all factors, including the prosumer behavior, and has perfect foresight of future information. This assumption is not practical in real-world scenarios. Despite being influenced by the randomness of the prosumer behavior, the DRL-avg performance can still be improved by increasing the battery capacity.

In conclusion, increasing battery capacity can improve the duck curve. It is important to determine the optimal battery system capacity while considering the initial cost of batteries and the potential revenue of the RA.

G. Time Scalability

In this section, we compare the computational time using the baseline optimization method and the proposed DRL strategy with the different numbers of prosumers. Table IV summarizes the average CPU time at execution (online), i.e., calculating the solution, and the training time to converge the policy (offline). Regarding the online execution time, the optimization-based method increases the computational time exponentially with the number of prosumers, which is not practical due to the online time limitation ($\Delta t = 30min$). Meanwhile, the proposed DRL strategy takes only a few milliseconds, and it does not scale to the number of prosumers. The offline training time of the proposed DRL strategy remains in the practical range.

V. Conclusion

In this paper, we investigated the strategy of the resource aggregator (RA) to improve the duck curve. The DP-BS problem aims to maximize the RA's profit, minimize the prosumer's cost, and maximize the improvement of the duck curve. First, we have formulated the DP-BS problem as an MDP for a hierarchical energy market model. Then a model-free DRL algorithm has been used to learn the optimal strategy, which determines the retail price of prosumers and the charging/discharging of the battery power station. No prior knowledge of the details of prosumer and wholesale electricity prices is required to learn the strategy by using the proposed method. Therefore, the proposed method not only addresses the uncertainty of the system but also protects the privacy of the prosumers. The simulation results show that the proposed method with the best reward function can reduce the standard variation and the PAR of the net load by up to 57.1% and 23%, respectively, compared to baselines.

In our future work, we will concentrate on interactive weight selection for the multi-objective reward function based on entity's preferences, integrating the proposed approach with day-ahead scheduling, and detailed power system modeling to account for network constraints. Furthermore, while this paper primarily focuses on load shifting to later time slots, we recognize that load shifting to earlier time slots can still occur and may offer benefits in addressing the duck curve. However, as optimizing retail prices for load shifting to earlier time slots requires the RA agent to have knowledge of potential future elastic loads, the current model-free DRL-based framework may not perform optimally in such scenarios and may not converge well. Therefore, another future work is to develop a dynamic pricing framework that accommodates load shifting to both earlier and later time slots. This can be achieved by incorporating load forecasting of potential future elastic loads into our DRL-based approach.

References

[1] O. Gandhi, D. S. Kumar, C. D. Rodríguez-Gallegos, and D. Srinivasan, "Review of power system impacts at high PV penetration—Part I: Factors limiting PV penetration," *Solar Energy*, vol. 210, pp. 181–201, Nov. 2020.

[2] Y. Li, W. Gao, and Y. Ruan, "Performance investigation of grid-connected residential PV-battery system focusing on enhancing self-consumption and peak shaving in Kyushu, Japan," *Renew. Energy*, vol. 127, pp. 514–523, Nov. 2018.

[3] P. Olczak, P. Jaśko, D. Kryzia, D. Matuszewska, M. I. Fyk, and A. Dyczko, "Analyses of duck curve phenomena potential in polish PV prosumer households' installations," *Energy Rep.*, vol. 7, pp. 4609–4622, Nov. 2021.

[4] H. Zhong, Z. Tan, Y. He, L. Xie, and C. Kang, "Implications of COVID-19 for the electricity industry: A comprehensive review," *CSEE J. Power Energy Syst.*, vol. 6, no. 3, pp. 489–495, Sep. 2020.

[5] Q. Hou, N. Zhang, E. Du, M. Miao, F. Peng, and C. Kang, "Probabilistic duck curve in high PV penetration power system: Concept, modeling, and empirical analysis in China," *Appl. Energy*, vol. 242, pp. 205–215, May 2019.

[6] M. Doroshenko, S. Keshav, and C. Rosenberg, "Flattening the duck curve using grid-friendly solar panel orientation," in *Proc. e-Energy*, Jun. 2018, pp. 375–377.

[7] J. Meus, K. Poncelet, and E. Delarue, "Applicability of a clustered unit commitment model in power system modeling," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 2195–2204, Mar. 2018.

[8] *Renewable Energy Statics 2021*, Int. Renew. Energy Agency, Masdar City, UAE, 2021.

[9] R. Torabi, A. Gomes, and F. Morgado-Dias, "The duck curve characteristic and storage requirements for greening the island of Porto Santo," in *Proc. ES2DE*, Jul. 2018, pp. 1–7.

[10] M. Sheha, K. Mohammadi, and K. Powell, "Solving the duck curve in a smart grid environment using a non-cooperative game theory and dynamic pricing profiles," *Energy Convers. Manag.*, vol. 220, Sep. 2020, Art. no. 113102.

[11] M. Q. Raza, M. Nadarajah, and C. Ekanayake, "On recent advances in PV output power forecast," *Sol. Energy*, vol. 136, pp. 125–144, Oct. 2016.

[12] D. Watari et al., "Multi-time scale energy management framework for smart PV systems mixing fast and slow dynamics," *Appl. Energy*, vol. 289, May 2021, Art. no. 116671.

[13] B. Li, J. Shen, X. Wang, and C. Jiang, "From controllable loads to generalized demand-side resources: A review on developments of demand-side resources," *Renew. Sustain. Energy Rev.*, vol. 53, pp. 936–944, Jan. 2016.

[14] Energy Information Administration (EIA). "Electric power monthly—U.S." Accessed: Sep. 11, 2022. [Online]. Available: https://www.eia.gov/electricity/monthly/

[15] L. Jia and L. Tong, "Dynamic pricing and distributed energy management for demand response," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 1128–1136, Mar. 2016.

[16] A. Asadinejad and K. Tomsovic, "Optimal use of incentive and price based demand response to reduce costs and price volatility," *Elect. Power Syst. Res.*, vol. 144, pp. 215–223, Mar. 2017.

[17] A. Faruqui, "The ethics of dynamic pricing," *Elect. J.*, vol. 23, no. 6, pp. 13–27, Jul. 2010.

[18] S. Kerscher and P. Arboleya, "The key role of aggregators in the energy transition under the latest European regulatory framework," *Int. J. Elect. Power Energy Syst.*, vol. 134, Jan. 2022, Art. no. 107361.

[19] X. Lu, K. Li, H. Xu, F. Wang, and Y. Zhang, "Fundamentals and business model for resource aggregator of demand response in electricity markets," *Energy*, vol. 204, May 2020, Art. no. 117885.

[20] K. Liu, Q. Chen, C. Kang, W. Su, and G. Zhong, "Optimal operation strategy for distributed battery aggregator providing energy and ancillary services," *J. Mod. Power Syst. Clean Energy*, vol. 6, no. 4, pp. 722–732, Jul. 2018.

[21] J. Jiang, Y. Kou, Z. Bie, and G. Li, "Optimal real-time pricing of electricity based on demand response," *Energy Procedia*, vol. 159, pp. 304–308, Feb. 2019.

[22] Z. Wang, M. Sun, C. Gao, X. Wang, and B. C. Ampimah, "A new interactive real-time pricing mechanism of demand response based on an evaluation model," *Appl. Energy*, vol. 295, Aug. 2021, Art. no. 117052.

[23] Z. Yang, M. Ni, and H. Liu, "Pricing strategy of multi-energy provider considering integrated demand response," *IEEE Access*, vol. 8, pp. 149041–149051, 2020.

[24] H. Taherian, M. R. Aghaebrahimi, L. Baringo, and S. R. Goldani, "Optimal dynamic pricing for an electricity retailer in the price-responsive environment of smart grid," *Int. J. Elect. Power Energy Syst.*, vol. 130, Sep. 2021, Art. no. 107004.

[25] R. Jovanovic, S. Bayhan, and I. S. Bayram, "A multiobjective analysis of the potential of scheduling electrical vehicle charging for flattening the duck curve," *J. Comput. Sci.*, vol. 48, Jan. 2021, Art. no. 101262.

[26] H. O. R. Howlader, M. M. Sediqi, A. M. Ibrahimi, and T. Senjyu, "Optimal thermal unit commitment for solving duck curve problem by introducing CSP, PSH and demand response," *IEEE Access*, vol. 6, pp. 4834–4844, 2018.

[27] I. Calero, C. A. Cañizares, K. Bhattacharya, and R. Baldick, "Duck-curve mitigation in power grids with high penetration of PV generation," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 314–329, Jan. 2022.

[28] A.-Y. Yoon, H.-K. Kang, and S.-I. Moon, "Optimal price based demand response of HVAC systems in commercial buildings considering peak load reduction," *Energies*, vol. 13, no. 4, p. 862, Feb. 2020.

[29] J. Ferdous et al., "Optimal dynamic pricing for trading-off user utility and operator profit in smart grid," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 2, pp. 455–467, Feb. 2020.

[30] J. Zhang, L. Che, L. Wang, and U. K. Madawala, "Game-theory based V2G coordination strategy for providing ramping flexibility in power systems," *Energies*, vol. 13, no. 19, p. 5008, Sep. 2020.

[31] M. Askeland, T. Burandt, and S. A. Gabriel, "A stochastic MPEC approach for grid tariff design with demand-side flexibility," *Energy Syst.*, to be published.

[32] A. S. Farsangi, S. Hadayeghparast, M. Mehdinejad, and H. Shayanfar, "A novel stochastic energy management of a microgrid with various types of distributed energy resources in presence of demand response programs," *Energy*, vol. 160, pp. 257–274, Oct. 2018.

[33] S. Abapour, B. Mohammadi-Ivatloo, and M. T. Hagh, "Robust bidding strategy for demand response aggregators in electricity market based on game theory," *J. Clean. Prod.*, vol. 243, Jan. 2020, Art. no. 118393.

[34] Y. Li, K. Wang, B. Gao, B. Zhang, X. Liu, and C. Chen, "Interval optimization based operational strategy of integrated energy system under renewable energy resources and loads uncertainties," *Int. J. Energy Res.*, vol. 45, no. 2, pp. 3142–3156, Feb. 2021.

[35] R. S. Sutton and A. G. Barto, *Reinforcement Learning, Second Edition: An Introduction*. Cambridge, MA, USA: MIT Press, Nov. 2018.

[36] B. Wang, Y. Li, W. Ming, and S. Wang, "Deep reinforcement learning method for demand response management of interruptible load," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3146–3155, Jul. 2020.

[37] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "A deep reinforcement learning method for pricing electric vehicles with discrete charging levels," *IEEE Trans. Ind. Appl.*, vol. 56, no. 5, pp. 5901–5912, Sep./Oct. 2020.

[38] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Appl. Energy*, vol. 220, pp. 220–230, Jun. 2018.

[39] Y. Kuang, X. Wang, H. Zhao, T. Qian, J. Wang, and X. Wang, "Model-free demand response scheduling strategy for virtual power plant considering risk attitude of consumer," *CSEE J. Power Energy Syst.*, to be published.

[40] D. Watari, I. Taniguchi, and T. Onoye, "Improving duck curve by dynamic pricing and battery scheduling based on a deep reinforcement learning approach," in *Proc. BuildSys*, Nov. 2021, pp. 232–233.

[41] N. Sadeghianpourhamami, T. Demeester, D. F. Benoit, M. Strobbe, and C. Develder, "Modeling and analysis of residential flexibility: Timing of white good usage," *Appl. Energy*, vol. 179, pp. 790–805, Oct. 2016.

[42] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," *Sustain. Energy Grids Netw.*, vol. 25, Mar. 2021, Art. no. 100413.

[43] P. R. Thimmapuram and J. Kim, "Consumers' price elasticity of demand modeling with economic effects on electricity markets using an agent-based model," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 390–397, Mar. 2013.

[44] M. Miller and A. Alberini, "Sensitivity of price elasticity of demand to aggregation, unobserved heterogeneity, price trends, and price endogeneity: Evidence from U.S. data," *Energy Policy*, vol. 97, pp. 235–249, Oct. 2016.

[45] P. Yang, G. Tang, and A. Nehorai, "A game-theoretic approach for optimal time-of-use electricity pricing," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 884–892, May 2013.

[46] P. Samadi, A.-H. Mohsenian-Rad, R. Schober, V. W. S. Wong, and J. Jatskevich, "Optimal real-time pricing algorithm based on utility Maximization for smart grid," in *Proc. SmartGridComm*, Oct. 2010, pp. 415–420.

[47] K. M. Tsui and S. C. Chan, "Demand response optimization for smart home scheduling under real-time pricing," *IEEE Trans. Smart Grid*, vol. 3, no. 4, pp. 1812–1821, Dec. 2012.

[48] M. S. Javadi, M. Gough, M. Lotfi, A. E. Nezhad, S. F. Santos, and J. P. S. Catalão, "Optimal self-scheduling of home energy management system in the presence of photovoltaic power generation and batteries," *Energy*, vol. 210, Nov. 2020, Art. no. 118568.

[49] Y. Kim, S. Leyffer, and T. Munson, "MPEC methods for bilevel optimization problems," in *Bilevel Optimization: Advances and Next Challenges*, S. Dempe and A. Zemkoho, Eds. Cham, Switzerland: Springer, 2020, pp. 335–360.

[50] M. van Otterlo and M. Wiering, "Reinforcement learning and Markov decision processes," in *Reinforcement Learning: State-of-the-Art*, M. Wiering and M. van Otterlo, Eds. Berlin, Germany: Springer, 2012, pp. 3–42.

[51] L. Han, Y. Peng, Y. Li, B. Yong, Q. Zhou, and L. Shu, "Enhanced deep networks for short-term and medium-term load forecasting," *IEEE Access*, vol. 7, pp. 4045–4055, 2019.

[52] M. Tanaka, "Real-time pricing with ramping costs: A new approach to managing a steep change in electricity demand," *Energy Policy*, vol. 34, no. 18, pp. 3634–3643, Dec. 2006.

[53] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[54] California ISO. "Wholesale electricity prices." Accessed: Jan. 20, 2021. [Online]. Available: http://www.caiso.com/Pages/default.aspx

[55] C. Miller et al., "The building data genome project 2, energy meter data from the ASHRAE great energy predictor III competition," *Sci. Data*, vol. 7, no. 1, p. 368, Oct. 2020.

[56] California Solar Initiative (CSI). "California Distributed Generation Statistics." Accessed: Jan. 20, 2022. [Online]. Available: https://www.californiadgstats.ca.gov/downloads/

[57] G. Brockman et al., "OpenAI gym," 2016, *arXiv:1606.01540*.

[58] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021.

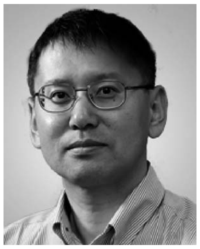
[59] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation Hyperparameter optimization framework," in *Proc. KDD*, 2019, pp. 2623–2631.

[60] L. Engstrom et al., "Implementation matters in deep policy gradients: A case study on PPO and TRPO," May 2020, *arXiv:2005.12729*.

[61] M. Andrychowicz et al., "What matters in on-policy reinforcement learning? A large-scale empirical study," Jun. 2020, *arXiv:2006.05990*.

Daichi Watari (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Osaka University, Japan, in 2018, 2020, and 2023, respectively, where he is an Alumnus with the Graduate School of Information Science and Technology. He was receiving support from JSPS Research Fellowship for Young Scientists (DC2). His research interests include decision-making using mathematical optimization and reinforcement learning for an energy management system.

Ittetsu Taniguchi (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Osaka University in 2004, 2006, and 2009, respectively. From 2007 to 2008, he was an International Scholar with Katholieke Universiteit Leuven (IMEC), Belgium. In 2009, he joined the College of Science and Engineering, Ritsumeikan University as an Assistant Professor, and became a Lecturer in 2014. In 2017, he joined the Graduate School of Information Science and Technology, Osaka University as an Associate Professor. His research interests include system-level design methodology and design methodologies for cyber–physical systems. He is a member of ACM, IEICE, and IPSJ.



Takao Onoye (Senior Member, IEEE) received the B.E. and M.E. degrees in electronic engineering and the Dr.Eng. degree in information systems engineering from Osaka University, Japan, in 1991, 1993, and 1997, respectively, where he is currently a Professor and the Dean of the Graduate School of Information Science and Technology, Osaka University. His research interests include media-centric low-power system architecture and implementation. He has also taken various volunteer positions in academic societies, such as an Editor-in-Chief of *IEICE Transactions on Fundamentals* (Japanese Edition), IEEE CAS Society Board of Governors, IEEE Region 10 Treasurer, IEEE Region 10 Vice Chair, and IEEE Japan Council Chair. He is a member of IEICE, IPSJ, and ITE-J.