A Strategic Day-ahead Bidding Strategy and Operation for Battery Energy Storage System by Reinforcement Learning

DOI:
10.1016/j.epsr.2021.107229

Document Version
Accepted author manuscript

A Strategic Day-ahead Bidding Strategy and Operation for Battery Energy Storage System by Reinforcement Learning

Yi Dong[a], Zhen Dong[a], Tianqiao Zhao[b], Zhengtao Ding[a,*]

[a]*Department of Electrical and Electronic Engineering, the University of Manchester, M13 9PL, Manchester, UK*
[b]*Department of Electrical and Computer engineering, Southern Methodist University, PO Box 750100, Dallas, TX 75275, USA*

Abstract

The Battery Energy Storage System (BESS) plays an essential role in the smart grid, and the ancillary market offers a high revenue. It is important for BESS owners to maximise their profit by deciding how to balance between the different offers and bidding with the rivals. Therefore, this paper formulates the BESS bidding problem as a Markov Decision Process (MDP) to maximise the total profit from the Automation Generation Control (AGC) market and the energy market, considering the factors such as charging/discharging losses and the lifetime of the BESS. In the proposed algorithm, function approximation technology is introduced to handle the continuous massive bidding scales and avoid the dimension curse. As a model-free approach, the proposed algorithm can learn from the stochastic and dynamic environment of a power market, so as to help the BESS owners to decide their bidding and operational schedules profitably. Several case studies illustrate the effectiveness and validity of the proposed algorithm.

Keywords: Battery energy storage system (BESS), power market bidding, reinforcement learning

Nomenclature

Indices and sets

$\mathcal{A}$	set of action variables
$\mathcal{M}$	set of Markovian decision processes
$\mathcal{P}$	set of transfer probabilities
$\mathcal{R}$	set of reward variables
$\mathcal{S}$	set of state variables
$charge$	subscript of BESS charge
$discha$	subscript of BESS discharge
$down$	superscript of regulation-down market
e	superscript of energy market
reg	superscript of regulation market
up	superscript of regulation-up market

Variables

η_c, η_d	charging and discharging efficiency
Profit	profit of a BESS
a	decision action variable
b_p	bidding price
b_c	regulation capacity bids
C^W	penalty function of operation constraints
C^{total}	total cost of a BESS owner
C_{ag}	ageing cost function of BESS
C_{loss}	bidding price
$C_{M\&O}$	maintenance and operation cost function
d	depth of discharge
E	energy state of BESS
N_d^{fail}	maximum number of charging/discharging cycles at d DoD
P	power output of a BESS
p	market clearing price
Q	value function related to state-action pair
r	immediate reward signal from environment
s	observation variable from environment
soc	state of charge
V	value function related to state variable
v^{-1}	clearing price of the previous day

Parameters

α, β	learning rate of reinforcement learning
ΔT	regulation period
γ	discount factor of reinforcement learning
$\mathcal{E}$	probability of exploration
ρ^{reg}	maximum ratio of regulation capacity
ρ_{max}	maximum efficiency opration rate of BESS
ρ_{min}	minimum efficiency opration rate of BESS
τ	time index of regulation step
ξ	component replacement cost of BESS
a, b	coefficients of charge and discharge
C_a	annual cost of BESS
C_E	Unit cost of energy storage system
C_F	Unit cost of facility infrastructure
C_P	Unit cost of power conversion system
C_{inv}	investment cost of BESS
E_{max}	maximum energy capacity of BESS
h	operation period
h_e	operation period in energy market
h_{reg}	operation period in regulation market
K_p	parameter for different types of BESS

*Corresponding author
E-mail address: zhengtao.ding@manchester.ac.uk (Z. Ding).

N_μ	number of maximum episodes
N_f	number of features
N_t	number of maximum steps
P_{max}	maximum power output of BESS
t	hourly time index
w_e	weight of energy market
w_{reg}	weight of regulation market

1. Introduction

Battery Energy Storage System (BESS) gets the opportunity to play an important role in the future smart grid. With the rapid development of battery technology, the BESS can bring more benefits for the owners and the cost of BESS construction is gradually reduced [1–3]. There will be more companies focusing on the development and construction of the BESS. As the BESS capacity increases, the BESS will participate in different markets and benefit from multiple services [4, 5]. Additionally, the frequency regulation market demands rapid response and offers high returns [6, 7], so that the BESS owners will put more attention on the regulation market with their BESS, which will lead to competition in the future smart grid. Therefore, how to allocate the capacity of BESS and make bidding decisions has become an important issue.

One major application for the BESS is frequency regulation services in the Automation Generation Control (AGC) market. BESS has the characteristics of easy storage, high reliability and fast response, which is more suitable than pumped-storage plant and heat storage plant for the frequency regulation market. Moreover, the AGC market offers 3 times mileages for Dynamic Regulation Signal D (RegD) service, which will bring high revenue for the BESS owners. As a result, more BESS owners are expected to compete in the AGC market and some researchers have been paying more attention to the AGC market [8–13]. In [8], a control strategy for the BESS in frequency regulation was provided, considering the ageing cost while keeping the State of Charge (SoC) of the BESS. In [10], a coordinated control strategy of BESS was proposed to ensure the wind power plants' commitment to frequency ancillary services, focusing on reducing the BESS's size and extending the lifetime of the BESS. However, mentioned literature only consider the application of the BESS in one market. With the emergence of large-capacity BESS, some articles study the operation strategies of the BESS in multiple markets, so as to maximise the overall profit of the BESS by controlling the placement proportion of the BESS in different markets. For example, He, et al. [12] integrated the energy storage system and solar power plant and proposed an optimal strategy for Concentrating Solar Power (CSP) plant, which considered the energy, reserve and regulation market. He also proposed a Performance-Based Regulation (PBR) based optimal bidding model in [13]. It not only addressed the optimal strategy for the BESS in different markets but also considered the battery life.

Another problem missed by these literature is that the bidding strategies only solve the allocation problem of the single BESS, in which their bidding rivals are neglected. With the entry of the rivals, the bidding market of the BESS presents some challenges [14]. During the process of bidding, the bidder does not know the rivals' bidding price and bidding quantity, which is hard to solve by traditional optimisation algorithms. Furthermore, since bidding is a highly random and uncertain process, the bidders cannot know the specific revenue model during bidding. They only know the offer results from the System Operator (SO) in the smart grid. Considering incomplete information of stochastic demand from the market and unknown bids from rivals, some individual based approaches have been widely applied for bidding strategies in electric market, where the individual agent learns to maximise its own profit [15–17]. For example, Kebriaei, et al. [18] combined state estimation and fuzzy Q-learning to learn the optimal decision of the generators. Li, et al. [19] applied the model-free reinforcement learning algorithm to solve the optimal carbon capture in the wholesale market bidding problem. Nanduri, et al. [20] formulated a stochastic game model for the energy market and proposed a reinforcement learning based solution methodology. Lakic, et al. [21] simulated the market as a stochastic environment and proposed a novel agent based SA-Q-learning for demand-side system reserve provision. However, there is very little understanding of the potential benefits of BESS in the wider power system or micro-grids [22].

Therefore, this paper proposes a novel Markovian based bidding model that decides the optimised bidding strategy of the BESS in day-ahead energy and regulation markets, considering the charging/discharging losses and the ageing cost of the BESS. Additionally, the Function Approximation based Reinforcement Learning (FARL) algorithm is applied to the proposed model to solve the multiple rival bidding problem. The function approximation approach is introduced in this paper to address the redundancy caused by massive data, and therefore prevent the dimension curse. Based on the proposed model, the BESS could obtain a more accurate and profitable bidding strategy.

The major contributions of this paper are summarised as follows:

1. The BESS bidding problem is modelled as an MDP framework for learning the optimised bidding policy to increase the welfare of BESS in energy and regulation markets. The model has been delicately designed, especially considering the losses during the power transfer and the ageing cost of the BESS.
2. Since reinforcement learning (RL) involves discrete-state transition, a function approximation approach is introduced to transfer the uncertain and continuous bidding environment into a set of discrete states, such that the memory and computational complexities can be reduced. This makes the state transition tractable, and avoids the curse of dimensionality.
3. The proposed bidding strategy of BESS owners considers both energy market and regulation market, which shows flexibility to the uncertain bidding environments, such as prior knowledge of other rivals and dynamics of the sys-

tem operator. As an individual profit maximisation bidding strategy, it can help the BESS owner optimise its bidding strategy to obtain highest bidding revenue without rivals' information.

The rest of the paper is organised as follows. The market framework implemented in this paper is summarised in section 3. In section 4, a model of the BESS is formulated and the constraints are designed. In section 5, the FARL algorithm is introduced to find the bidding strategy for the BESS. Simulation results and corresponding analysis are presented in section 6. Finally, conclusions are drawn at the end of the paper.

2. Preliminaries

In this section, we recall some necessary concepts related to the reinforcement learning algorithm.

The reinforcement learning problems can be viewed as MDP, which is the stochastically changing system. It is composed of state $\mathcal{S}$, action $\mathcal{A}$, transition probability function $\mathcal{P}$, reward function $\mathcal{R}$ and discount factor γ. Therefore, MDP is defined as a five-element tuple in this paper:

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma\} \tag{1}$$

At each time slot t, the intelligent agent has its observation of the environment, namely state s_t. Then, the agent will choose its action followed by a policy function $\pi(s_t) : \mathcal{S} \to \mathcal{A}$, which denotes a distribution over actions for each state. In the reinforcement learning, there is a transition function $\mathcal{P}(s_t, a_t, s_{t+1}) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. It maps state s_t to s_{t+1} by action a_t, which means the dynamics of the environment. The transition function is unknown and has part of stochastic factors. Thus, the agent needs to learn it through different $\{s_t, a_t, s_{t+1}\}$ sets during the training process. Specific to each state transition between adjacent time slots, the environment will provide a reward signal $r_t \in \mathcal{R}$ to the agent. Then the trajectory $\{s_0, a_0, s_1, a_1, \cdots, s_T, a_T\}$ can be derived with the discounted trajectory return $\sum_{t=1}^{T} \gamma^t R(s_t, a_t)$. For any policy π, the value function of state s can be defined as the expected total discounted reward:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t R(s_t, a_t) | s_t = s\right], \forall s \in \mathcal{S} \tag{2}$$

Then the corresponding state-action value function Q^{π} is defined as:

$$Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t R(s_t, a_t) | s_t = s, a_t = a\right], \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{3}$$

According to the Bellman equation [23], the value function Q^{π} can be represented in a recursive format:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\left[R(s_t, a_t, s_{t+1}) + \gamma Q^{\pi}(s_{t+1}, \pi(s_{t+1}))\right] \tag{4}$$

where $R(s_t, a_t, s_{t+1})$ is the observed reward after taking action a_t at state s_t and resulting in state s_{t+1}. The equation (4) indicates that the Q function can be improved by using current value of the Q^{π} estimation. To reduce computational complexity, Temporal Difference (TD) learning is one of the most famous update methods instead of Monte-Carlo and Dynamic Programming. It only requires current state s_t, current action a_t, reward $R(s_t, a_t, s_{t+1})$ and next state s_{t+1}:

$$Q^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) + \alpha \delta_t \tag{5}$$

$$\delta_t = r_t + \gamma \max_{a_{t+1}} Q^{\pi}(s_{t+1}, a_{t+1}) - Q^{\pi}(s_t, a_t) \tag{6}$$

where $0 \leq \alpha \leq 1$ is the learning rate. δ_t is the TD error at the time slot t, which implies the correction between the estimation and target value of Q function. When time goes to infinity, $Q^{\pi}(s, a)$ will converge to its optimal value $Q^*(s, a)$ for all state-action pairs. Here, the optimal value function $Q^*(s, a) = \sup_{\pi} Q^{\pi}(s, a)$ is defined for all state action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$. With the optimal Q^* function, the optimal policy π^* can be obtained by greedy algorithm:

$$\pi^*(a|s) = 1, \quad \text{if} \quad Q^{\pi}(s, a) = \max_{a'} Q^*(s, a') \tag{7}$$

where a' is any possible action associate with state s.

3. Market Design

This section studies the bidding mechanism of battery energy storage system in different power markets. In this paper, we assume that the BESS can offer more than one service in different markets. The BESS owner has to provide the day-ahead hourly bids to the system operator, including bidding capacities and bidding prices. The system operator determines the requirements of different services according to short-term load forecasting, renewable energy prediction and reliability constraints. On the basis of these, the Market Clearing Price (MCP) and offers of different markets are derived related to the different quotations and capacity bids. During the bidding process, the participants cannot know the bidding data of their rivals, but the MCP and offers from the system operator are public.

With the development of battery technology, the capacity of BESS is increasing rapidly. According to the importance of batteries in AGC market service, we assume that the BESSs have the market power to influence AGC market [2]. Since the main services and revenues of BESS come from the AGC market, according to [5], supplying sufficient power and energy capacity for the AGC market has the highest priority among all the services from the perspective of the system operator. In this paper, based on the prediction of energy market and AGC market, the winning bids of BESS are determined considering the AGC market conditions.

3.1. Automatic Generation Control (AGC) Market

In the AGC market, the operation of smart grid must be subjected to keep the supply and demand balance. During the frequency control, the supply-demand balance of the whole network is met by adjusting the output of frequency modulation

units, such as BESS, capacitive energy storage system, superconducting magnetic energy storage system, thermal energy storage system and Flywheel energy storage system [24]. In the frequency adjustment, there are various components involved, shown in Fig. 1.

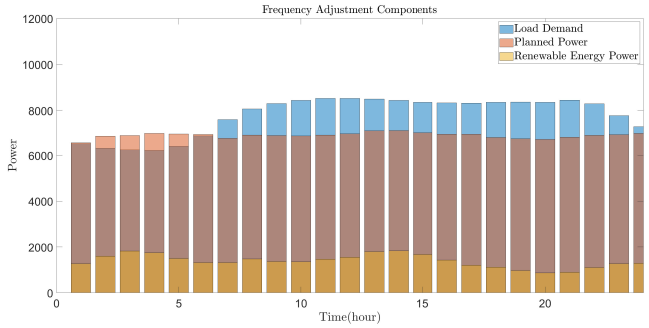

Figure 1: Frequency adjustment components.

Therefore, the AGC market should take the whole grid demand as the benchmark, and obtain the mismatched power by calculating the capacity demand caused by load change, the renewable energy output power, and the planned output power.

$$\Delta P_{td} = P_{load} - P_{energy} - P_{plan} \tag{8}$$

where ΔP_{td} is the power mismatch; P_{load}, P_{energy} and P_{plan} are the load demand, renewable power output and the planned power output, respectively.

In power grid dispatching, Area Control Error (ACE) are usually sent to AGC with a period of 2-4 seconds.

$$P_{ACE} = \Delta P_{td} + \beta_f \cdot \Delta f \tag{9}$$

where β_f is the coefficient of frequency deviation, Δf is the frequency deviation and P_{ACE} is the ACE signal.

From 2017, the conditional neutrality controller has been applied to control the regulation resources in PJM market [25]. It is a hybrid PID controller which includes a RegD integral feedback loop to ensure the energy of RegD is neutral. If system conditions allow, RegA will be utilised to balance the neutrality of RegD. For example, if the RegA resources are fully utilised to control ACE, then it will not be able to assist RegD. PJM area control error (ACE) signal is fed to high-pass/low-pass filters and a PID regulation controller to generate two regulation signals: a fast responding dynamic regulation signal D (RegD) and a slow responding traditional regulation signal A (RegA) [25, 26].

Frequency regulation mileage refers to the sum of the absolute changes in output power within a period of time, and it is usually measured in megawatts (MW). As shown in Fig. 2, the RegA signal moves much slower than RegD signal. The mileage of RegD is about 7.17 times that of RegA in PJM market in 2019 [27]. Therefore, RegD offers more opportunity and higher performance compensation to exploit the potential of fast response energy storage systems. The RegD signal changes every 2 to 4 seconds, and the response time of BESS is usually on the time scale of seconds or milliseconds. Nevertheless,

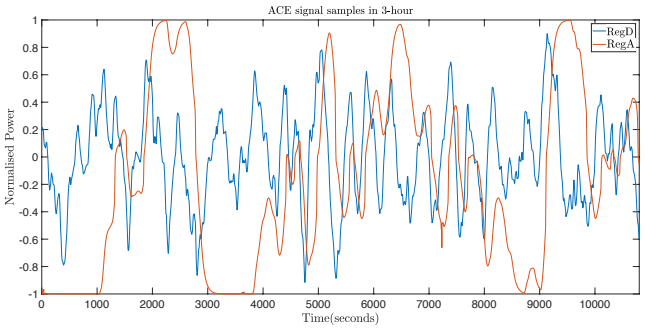

Figure 2: Real-time RegD and RegA data.

PJM market requires the mean value of RegD signal to be zero, which is suitable for energy limited power units like BESS.

3.2. Energy Market

In our model, the revenue of energy market is mainly from the planned output power. Compared with traditional generating units, a BESS only supplies or consumes small portion of electricity, the BESSs are supposed to be the price-takers, who will not affect the energy price in the energy market. The BESS will submit the day-ahead bids to the energy market system operator, and then the system operator will allocate the electric energy according to different requirements. Since BESS has the characteristics of low cost, good power quality and fast response, we assume that the battery will win the bids in the energy market. Therefore, the revenue in the energy market can be described as:

$$R_{e,t} = p_t^e \cdot b_{e,t} \tag{10}$$

where p_t^e is the electricity price in energy market, $b_{e,t}$ is the energy bidding quantity of the BESS and $R_{e,t}$ is the revenue of BESS in energy market at time slot t.

3.3. Model of BESS

The BESS unit should provide AGC services frequently in long term running. Therefore, two types of BESS costs are considered in this paper, i.e., charging/discharging loss cost and the BESS ageing cost.

3.3.1. Loss Cost of BESS

According to [28], charging efficiency and discharge efficiency are different, and the charging/discharging efficiency can be formulated as η_c and η_d, respectively. We assume that the energy price is p_t^e. The charging/discharging losses then represented as

$$C_{chaloss} = p_t^e \cdot P_{charge}(1 - \eta_c) \cdot \Delta T \tag{11}$$

$$C_{disloss} = p_t^e \cdot P_{discha}(\frac{1}{\eta_d} - 1) \cdot \Delta T \tag{12}$$

where ΔT is the control period of regulation service and it is set as 4 seconds.

3.3.2. Ageing Cost of BESS

Ageing cost is an important expenditure when BESSs provide the power system service, and the BESS may not meet the requirements of the system after excessive ageing. Therefore, the ageing cost model needs to be considered when calculating the revenue of the BESS. Based on [29, 30], the maximum energy capacity of BESS will be reduced by the increase of charging/discharging cycles. In addition, the smaller the depth of discharge is, the more cycles there will be [31]. For a given depth of discharge of a lead-acid battery, the number of cycles before failure can be seen in Fig. 3 below.

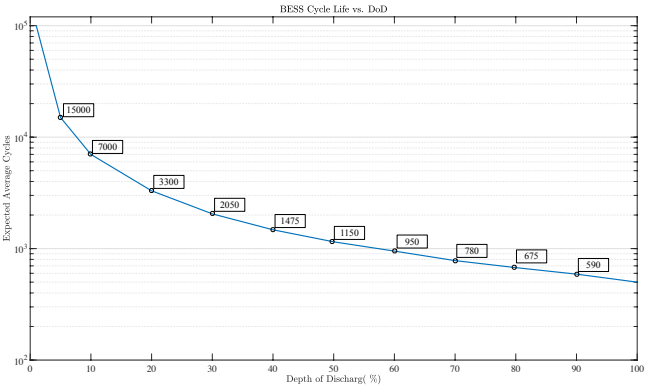

Figure 3: The relationship between DoD and cycle life of a BESS.

For different types of battery, N_d^{fail} is a function of DoD(%), which can be calculated as

$$N_d^{fail} = N_{100}^{fail} \cdot d^{(-k_P)} \tag{13}$$

where N_d^{fail} is the maximum number of charging/discharging cycles at d DoD, d is the depth of discharge (DoD), and k_P is a constant parameter for different types of batteries ranging from 1.1 to 2.2 [32]. In the reinforcement learning algorithm, the time interval between s_t and s_{t+1} is one hour. Depth of Discharge (DoD) is the fraction or percentage of the capacity which has been removed from the BESS, which can only be calculated after a charging or discharging event. However, the time interval of the regulation market publishing regulation signal is 2-4 seconds, which means that there will be one charging or discharging event in each 2-4 seconds interval. The ageing cost between s_t and s_{t+1} should be the sum of every charging/discharging event's ageing cost. The DoD of BESS is changed in each half cycle, so that d will response to each RegD signal in each time period $(\tau - \Delta T) \rightarrow \tau$. Meanwhile, the BESS can only accept one charging mission or one discharging mission at a time. Thus, we can formulate the ageing cost for one event, named half cycle, as $C_{ag,\tau}^{half}$, which can be calculated as:

$$C_{ag}^{half} = \frac{(d_\tau)^{k_P}}{2 \cdot N_{100}^{fail}} \cdot C_{inv} \tag{14}$$

where C_{inv} is the average daily investment cost of the battery energy storage system, which can be calculated by

$$C_{inv} = (1 + \xi) \cdot (C_P \cdot P_{max} + C_E \cdot E_{max} + C_F) \tag{15}$$

where ξ is the component replacement cost. $C_E \cdot E_{max}$ is the cost of the storage unit, where E_{max} is the energy capacity of the BESS. $C_P \cdot P_{max}$ is the costs of power conversion system, P_{max} is the power capacity of BESS. C_P, C_E and C_F are the unit costs of power conversion system, energy storage and facility infrastructure costs, respectively.

In the bidding market, the BESS company cannot predict the positive and negative power command signals to be given by the system operator, so that the BESS owner will provide one bid for charging and one bid for discharging at each time slot [12]. Therefore, we could get the equivalent one cycle ageing cost as:

$$C_{ag,\tau} = \frac{|(d_{\tau+1})^{k_P} - (d_\tau)^{k_P}|}{2 \cdot N_{100}^{fail}} \cdot C_{inv} \tag{16}$$

$$d_{\tau+1} = d_\tau + \frac{P \cdot \Delta T}{E} \tag{17}$$

Combining (13)-(17), we can obtain the equivalent ageing cost for one hour as:

$$C_{ag} = \sum_{\tau=1}^{3600/\Delta T} \frac{|(d_{\tau+1})^{k_P} - (d_\tau)^{k_P}|}{2 \cdot N_{100}^{fail}} \cdot C_{inv} \tag{18}$$

where ΔT is the time step of frequency regulation signal, and there are $3600/\Delta T$ charging/discharging cycles within one hour.

4. Model Formulation

The proposed model of BESS bidding in the pool based electricity market is described in detail. The decision variables are the capacity bids in energy market $b_{e,t}$, the capacity bids in AGC market $b_{c,t}^{up}$ and $b_{c,t}^{down}$ and the price bids in AGC market $b_{p,t}$ of the BESS for each hour in the next day.

4.1. Objective Function

The bidding model is to maximise the total profit of a BESS owner, which is described as follows

$$\max \text{Profit} = \sum_{t \in T} (\text{Profit}_t^{\text{e}} + \text{Profit}_t^{\text{reg}} - \text{Cost}_t^{\text{total}}) \tag{19}$$

where $\text{Profit}_t^{\text{e}}$ and $\text{Profit}_t^{\text{reg}}$ are the hourly revenue from energy market and the regulation market, respectively. $\text{Cost}_t^{\text{total}}$ is the hourly cost, which includes operation and maintenance cost, charging/discharging cost and the ageing cost. t is the hour index and Profit_t is the 24-hour total profit.

In the electricity market, there is a system operator between the supply companies and the retailers. The suppliers are bidding in the power pools, and the system operator makes the decision of market price and power generation offers. Since the BESSs are the price-takers in the energy market, the total revenue of a BESS in energy market $\text{Profit}_t^{\text{e}}$ can be calculated by [33, 34].

$$\text{Profit}_t^{\text{e}} = p_t^e \cdot b_{e,t} \cdot h_e \tag{20}$$

$$P_{e,t} = \begin{cases} b_{e,t} \cdot \dfrac{1}{\eta_d}, & \text{if} \quad b_{e,t} > 0 \\ b_{e,t} \cdot \eta_c, & \text{if} \quad b_{e,t} < 0 \end{cases} \tag{21}$$

where $b_{e,t}$ is the winning power offer of the BESS at time slot t, termed as the capacity bidding quantity. h_e is the normally operation period in energy market, typically 1 hour or 15 mins. $P_{e,t}$ is the charged and discharged power in the BESS. Note that $b_{e,t}$ can be positive or negative, which is related to charging and discharging requirement. A power supplier can only generate power if its offers are accepted. Otherwise, the extra penalties should be paid. The subscript "t" is the index of the hours in each day, since the bidding strategy is day-ahead with hourly bids in the wholesale electricity market.

In (19), $\text{Profit}_t^{\text{reg}}$ is the revenue of the regulation markets, which can be described as

$$\text{Profit}_t^{\text{reg}} = \text{Profit}_t^{cap} + \text{Profit}_t^{perf} \tag{22}$$

where Profit_t^{cap} is the revenue of the regulation capability, which can be described as

$$\text{Profit}_t^{cap} = (P_{cap,t}^{up} + P_{cap,t}^{down}) \cdot p_t^{cap} \cdot h_{reg} \tag{23}$$

where $b_{c,t}^{up}$ and $b_{c,t}^{down}$ are the capacity bids; p_t^{cap} is the Regulation Market Capacity Clearing Price (RMCCP), which is influenced by the bidding prices. h_{reg} is the normally operation period in regulation market, and it is typically 1 hour and 15 mins. Different energy storage systems provide different regulation capacity bids. Then the system operator will make the decision and send the regulation signal to the frequency modulation units. If the regulation bid of the BESS is accepted by the system operator, the regulation capability compensation and the regulation performance based profit can be formulated as

$$\text{Profit}_t^{perf} = (P_{cap,t}^{up} + P_{cap,t}^{down}) \cdot p_t^{perf} \cdot s_c \cdot \Delta T \tag{24}$$

where p_t^{perf} is the Regulation Market Performance Clearing Price (RMPCP). s_c is the performance score, which related to the accuracy, delay and precision [7]. ΔT is the regulation period, typically from 2s - 4s. According to the report [7], the performance revenue is not related to the bidding capacity of the BESS, but the real-time regulation signal and the clearing price. Since each time slot, the regulation signal will only have one sign, we separate the regulation signal into regulation up signal $P_{cap,\tau}^{up}$ and the regulation down signal $P_{cap,\tau}^{down}$, where τ is the time index of regulation step.

The total cost is calculated in (25).

$$\text{Cost}_t^{\text{total}} = C_{O\&M,t} + C_{loss,t} + C_{ag,t} \tag{25}$$

where $C_{O\&M,t}, C_{loss,t}$ and $C_{ag,t}$ are the operation and maintenance cost, charging/discharging cost and the ageing cost, respectively. The operation and maintenance cost of BESS is usually a variable cost proportional to the size of BESS, which can be calculated as $C_{O\&M,t} = C_a \times E_{max}$, where C_a is the annual maintenance cost of BESS [35].

The charging/discharging cost is the sum of the charging part and the discharging part. In this model, the charging power P_{charge} is equal to the regulation down signal $P_{cap,\tau}^{down}$ and $P_{discha} = P_{cap,\tau}^{up}$. Therefore, the charging/discharging cost for each hour is formulated as

$$C_{loss,t} = \sum_{\tau=1}^{3600/\Delta T} p_t^e \cdot (P_{cap,\tau}^{down}(1-\eta_c) + P_{cap,\tau}^{up}(\frac{1}{\eta_d} - 1)) \cdot \Delta T \tag{26}$$

The last part of total cost is the ageing cost, which can be calculated by (18).

4.2. Constraints

4.2.1. Power Constraints

In this part, the capacity limits of the BESS are considered and formulated in (27)-(29) regarding market requirements, physical constraints and regulation constraints. The sum of the BESS bids must be kept within the maximum power of the BESS.

$$P_{e,t} + P_{c,t}^{up} \leq P_{max} \tag{27}$$

$$P_{e,t} - P_{c,t}^{down} \geq -P_{max} \tag{28}$$

where P_{max} is the maximum output power of the BESS. It is related to the type of the BESS.

Furthermore, the maximum regulation capacity has to be limited in a reasonable range, described in (29).

$$0 \leq P_{c,t}^{up}, P_{c,t}^{down} \leq \rho^{reg} \cdot P_{max} \tag{29}$$

where ρ^{reg} is the maximum ratio of regulation capacity to the high sustained limit.

To meet the transmission constraints in power system, the BESS is required to hold enough energy to response the system operator for dispatch or reserves [36]. Therefore, we consider that the BESS must maintain the output power level for at least h_e for energy market and h_{reg} for regulation market [13].

$$E_{max} \cdot soc_t \geq (b_{e,t} \cdot h_e + b_{c,t}^{up} \cdot h_{reg})/\eta_d \tag{30}$$

$$E_{max} \cdot soc_t \leq E_{max} + (b_{e,t} \cdot h_e - b_{c,t}^{down} \cdot h_{reg})\eta_c \tag{31}$$

4.2.2. Charging/Discharging Constraints

This part models the energy balance model of the BESS based on the physical constraints and the market requirement.

We assume that there is no energy loss during the charging/discharging process. The SoC of the BESS can be calculated as:

$$soc_t = soc_{t-1} + \Delta soc_t \tag{32}$$

where Δsoc_t indicates the amount of energy change between time $t-1$ and t, which is usually expressed in percentage (%). According to the energy selling and buying, the value of Δsoc_t can be positive and negative. For different types of BESS, the charging efficiency are different. Therefore, the charging/ discharging rate of the BESS (Δsoc_t) is expressed as

$$\Delta soc_t = (\Delta E_t^e + \Delta E_t^c)/E_{max} \tag{33}$$

$$\Delta E_t^e = P_{e,t} \cdot h_e \tag{34}$$

$$\Delta E_t^c = (\eta_c \cdot P_{cap,t}^{down} - \frac{1}{\eta_d} \cdot P_{cap,t}^{up}) \cdot h_{reg} \tag{35}$$

where $\Delta E_t^e, \Delta E_t^c$ represent the amount of energy change in energy market and frequency regulation market, respectively. Note that the energy loss formulated here will not influence the reward calculation. soc_t is used to calculate the next state of

the reinforcement learning algorithm, which is the actual state of BESS.

The BESS must keep its SoC within its energy capacity limits. According to [37], the BESS performs its best working characteristics between 20% - 80%. To get the best performance of the BESS, in this paper, the capacity limits is set as

$$\rho_{min} \cdot E_m \leq soc_t \cdot E_{max} \leq \rho_{max} \cdot E_m \forall t \in T \tag{36}$$

where ρ_{min} and ρ_{min} are the minimum and maximum efficiency operation rate. E_m is the rated energy capacity of the battery storage.

4.2.3. SoC Constraints

The initial and final SoC usually are set to be same during the optimization period, as described below. t_0 and t_{24} represent the begin and end of the day.

$$soc_{t_0} = soc_{t_{24}} \tag{37}$$

5. Algorithm Design

5.1. Function Approximation based Reinforcement Learning

Reinforcement learning is a valid approach to solve the decision problem for an unknown and uncertain environment. In this paper, we deploy reinforcement Q-learning to achieve the optimal bidding results. In the traditional Q learning, it needs to generate a complicated Q table, and the dimension of the Q table will fall into the dimension curse with the increase of actions and states. Function approximation is a valid method to solve the generalisation problem for the large dimension of state and action pairs in the reinforcement learning method. Therefore, to avoid such dimension curse in BESS bidding problem, we apply the function approximation method to approximate the Q value. Here, the linear approximator analysed in this paper is:

$$Q(s_t, a_t) = \sum_{j=1}^{n} \phi_j(s_t, a_t)\theta_j = \phi_t^T(s_t, a_t)\theta_t \tag{38}$$

where θ_t is the approximation parameter vector with n elements, and $\phi_t(s_t, a_t)$ is the feature vector, which is given by

$$\phi_t(s_t, a_t) = \{\phi_1(s_t, a_t), \phi_2(s_t, a_t), \cdots, \phi_{N_f}(s_t, a_t)\} \tag{39}$$

where $\phi_j(s_t, a_t)\Re$ are the basis functions, and N_f is the number of features. The details of function approximation are presented in Appendix.

5.2. Problem Reformulation

Aiming at the stochastic environment of power market, the optimal bidding problem in an stochastic environment is reformulated based on equation (1), which includes the state space $\mathcal{S}$, action space $\mathcal{A}$, transition probability function $\mathcal{P}$, reward function $\mathcal{R}$ and discount factor γ in detail.

At each time slot t, the BESS owner has its observation of the bidding market, namely state s_t. Considering the bidding quantity and bidding price of rivals are uncertain, the state of the BESS owner is set as:

$$s_t = (v_t^{-1}, a_{t-1}^T, soc_t, t)^T \tag{40}$$

where $s_t \in \mathcal{S}$ presents the observable information and v_t^{-1} is the clearing price of the previous day at time slot t. a_{t-1} is the last decided bidding actions, including bidding quantities and bidding prices. In the wholesale electricity market, each BESS owner only knows its own bidding quantity and price. The bidding data of the other bidders must be estimated by the previous bidding history. In this paper, the bidding quantities and prices of other rivals are presumed to be influenced by the market clearance price and the sold offer at time slot $t-1$. Some similar state-choosing methods are studied for electricity market in [34]. soc_t is the SoC of BESS at time slot t, which can be accurately estimated by battery management systems. Here, our objective is to maximise the BESS owner's profit within its bidding period, which is 24 hours of a day; therefore, time slot t is set as a part of state so that the decision maker can take different actions in different hours of the day-ahead bidding strategy.

The action $a_t \in \mathcal{A}$ consists of decision variables made by BESS owner. Since bidding environment are unpredictable, we only formulate the decision variables concerning the bids of own BESS units:

$$a_t = (b_{e,t}, b_{p,t}, b_{c,t}^{up}, b_{c,t}^{down})^T \tag{41}$$

where $(b_{e,t}, b_{p,t}, b_{c,t}^{up}, b_{c,t}^{down})$ are the capacity bids in energy market, price bids, up and down capacity bids in AGC market, respectively. Here, $\mathcal{A}$ is the discrete action set for all s_t. At each time slot t, the BESS owner will provide a bidding action a_t from action set $\mathcal{A}$.

In the reinforcement learning, there is a transition function $\mathcal{P}(s_t, a_t, s_{t+1}) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$. It maps state s_t to s_{t+1} by action a_t, which means the dynamics of the environment. In the electric bidding market, the transition function is unknown and depends on some stochastic factors, such as real-time load mismatch and uncontrollable renewable power generation. Thus, our agent needs to learn it through different $\{s_t, a_t, s_{t+1}\}$ sets during the training process. After taking the action a_t, the state of BESS will automatically transfer to next state $s_{t+1} = (v_{t+1}^{-1}, a_t^T, soc_{t+1}, t+1)^T$ based on the transition function $\mathcal{P}$.

Specific to each state transition between adjacent time slots, the system operator will provide an offer to the BESS owner, which indicates the reward signal $r_t \in \mathcal{R}$. The algorithm can be trained by the reward information to select best policy to achieve maximum reward. In this paper, the detailed reward definition is designed as follows:

$$r_t = \text{Profit}_t - C_t^W \tag{42}$$

where r_{t+1} is defined under the framework of the reinforcement learning and $Profit_t$ is the one hour profit of the total profit $Profit$ in equation (21). Generally, the reward at $t+1$ time slot is the BESS owner profit in terms of the state s_t and the action a_t at t time slot. C_t^W is set as a penalty term, which is

related to the local constraints, including battery and generator constraints. For example, the SoC of a BESS should be kept between 20% to 80% to obtain the efficiency operation [38]. If the action leads to these inefficiency areas, the reward of this state action pair should be negative and get corresponding penalty. In this paper, the BESS owner can get the finite-time horizon reward sequence as $\{s_t, a_t, r_t, s_{t+1}, \cdots, s_{t+N-1}, a_{t+N-1}, r_{t+N-1}, s_{t+N}\}$, which is an episode of bidding and operation. The parameter N is the trading period, which is set as 24 in this paper.

The objective of the reinforcement learning for the BESS owner i is to obtain the best 24-hour reward given by

$$\mathbb{E}(\sum_{t=1}^{24} \gamma^t r_t | s_0) \tag{43}$$

where s_0 is the initial state; r_t is the reward based on state-action pair at time slot t; γ is the discount factor which is applied to reduce the effect of future reward. In this 24-hour bidding environment, $\gamma = 1$.

According to (3), a Q function can be defined as follows:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=1}^{24} \gamma^t r_t | s_t = s\right], \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{44}$$

Following the updating rules (5), $Q^\pi(s,a)$ will converge to the optimal Q value $Q^*(s,a)$. To ensure the proposed algorithm can find the optimal policy which covering maximum state values, we applied the $\mathcal{E}$-greedy policy to keep the exploration behaviour, so that all exploratory actions have probability to be chosen during the training period. The policy is settled by following equations:

$$\widetilde{a} = \arg\max_a Q^\pi(s,a), \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \tag{45}$$

$$\pi(s,a) = \begin{cases} 1 - \mathcal{E} + \mathcal{E}/n_A, & \text{if } a = \widetilde{a} \\ \mathcal{E}/n_A, & \text{if } a \neq \widetilde{a} \end{cases} \tag{46}$$

where $\widetilde{a}$ is the greedy action which has the maximum Q value under current policy π, n_A is the number of possible actions in action set $\mathcal{A}$, and $0 \leq \mathcal{E} \leq 1$ is the probability of choosing any action in the action set $\mathcal{A}$. Therefore, we have a $(1 - \mathcal{E} + \mathcal{E}/n_A) \times 100\%$ chance of taking the greedy action $\widetilde{a}$, and a $(\mathcal{E}(n_A - 1)/n_A) \times 100\%$ chance of exploring new behaviours.

Since the states setting in the model is continuous and the dimension of the states is large, this paper applies the function approximation to solve the reinforcement learning problem. An off-policy model-free algorithm is implemented so as to find the BESS bidding strategy, which helps the BESS to get a higher profit during the trading period.

5.3. Algorithm Implementation

Based on the equation (39), we randomly initialise the parameter θ_0 to calculate the the Q value of each state-action pair. To get quicker convergence speed of reinforcement learning algorithm, we apply a correction term w_0 to adjust the update law, which is designed in Appendix. Our algorithm is developed to update these parameters and get the optimal Q value Q^*.

In our algorithm, each hour is seen as a time step within N_t and each day is an episode within N_μ, which means the training process has $N_\mu \cdot N_t$ steps in total. At each episode, the algorithm will start from a random state s_0. Then, the agent chooses its action a_t according to policy π, then its state s_t will transfer to s_{t+1} and get the reward r_t. With all these obtained information and equations (5, 6, 65, 66), the algorithm can calculate and update the parameters θ_t, ω_t. After several explorations and training loops, our Q value Q^π is roughly equivalent to the optimal Q value Q^*. The details of the proposed algorithm for BESS optimal bidding are summarised in Algorithm I. It is started from a policy π, learning rate α, β and discount coefficient γ.

Algorithm 1 Function Approximation based Reinforcement Learning Algorithm for Supply-side BESS Bidding

Require: Learning rate α, β, Policy π, Discount coefficient γ
Ensure: The bidding action $a_i(t)$ of BESS owner for next day's trading market
  Initialisation: θ_0, ω_0
  for every episode $\mu = 1$ **to** N_μ **do**
    Initialise s_0, choose a for state s_0 with the $\mathcal{E}$-greedy policy π
    for every time slot $t = 1$ **to** N_t **do**
      Calculate the feature vector ϕ_t of state s_0
      Take action a_t, obeying the policy π
      Get the reward r_t from the environment
      Calculate the TD error $\delta(t)$
      Estimate the next step feature vector $\hat{\phi}_t$
      Update the parameter θ_i, ω_i
      $t \leftarrow t + 1$
    end for
    $\mu \leftarrow \mu + 1$
  end for
  return 24-hour Action Sequence;

6. Case Study

In this section, consider an electricity market with 4 BESSs, and these four BESSs bid in the AGC market to get their rewards. The planning horizon is next day 24-hour bids. In each state, BESSs make their decision for next day bids and each bid has capacity bidding price and capacity bidding quantity. And during the bidding process, BESS1 does not know how other BESSs are going to bid. However, the BESS1 can get the history clearing price information of the electricity market.

We carry out numerical simulations to investigate the computational efficiency of the proposed reinforcement learning algorithm. All the cases are performed using MATLAB on the computer equipped with a core Intel Core i7-6700 CPU and 16GB of RAM. After the training process, the average executive time of the all cases is about 19.7 milliseconds, which is promising to meet with the requirement of real-time bidding in power systems.

6.1. Datasets

In our simulation study, the real world datasets are applied to illustrate the effectiveness of our model. A 4-s based RegD & RegA signal is generated based on real RegD signal data by PJM's data set.

6.2. Case Implementation

In the case studies, it is supposed that all of 4 BESSs can participate in the AGC market. In this market, BESS1 is assumed as our own BESS, which tends to maximise the profit of next 24-hour. The decision maker of the BESS1 will implement the function approximation based reinforcement learning algorithm, seeking the proper bidding price and bidding capacities. In the stochastic bidding market environment, the bidding strategies of all the other BESS and the bidding environment are unknown, which means that BESS1 only have its individual bidding data and MCP history data. Similar to [39], the rule of clearing price is simply the highest bid from all accepted regulation offers in this paper. After that all market participates submit their hourly bids to the SO, the SO needs to schedule the regulation offer and publish an MCP according to the real-time load demand. Then, the BESS owners can calculate their rewards and costs based on the regulation offer and MCP. For the objective function shown in Eq. (19), the initial time slot is set as $t = 1$ and the end time slot is set as $t = 24$.

The charging efficiency of the BESS is derived to be 0.868 [12]. Since the function approximation algorithm is applied in the paper, the state variables in Eq. (40) will not be aggregated into discrete levels. And the action variables in Eq. (41) are aggregated into discrete levels to get more accurate results. The action aggregate is achieved by follows.

The bidding prices are set in advance by the BESS and the AGC market, and the other three capacity bids are considered together, since when two of them are specific in a domain, the other one should be constrained in some specific values. For example, if the regulation up and down capacities are defined, then the capacity bids of energy market should be limited in a specific domain.

Based on the real-time price data from PJM. the bidding price is aggregated into 11 levels, and the capacity bids are seperated into different 11 levels. Thus the test model has 14,641 aggregated actions.

The relevant information and cost parameters of the BESSs are shown in Tabs. 1 - 2.

6.3. Training Performance

In this section, we analyse the convergence characteristic of the function approximation based reinforcement learning algorithm. Note that since the action variables have been aggregated into 14641 elements and the states have been approximated into 400 elements, we have 5856400 theta variables. Here, the convergence curves of part of theta parameters are shown in Fig. 4.

Since the value of theta converges, the optimal Q-value of each state-action pair will automatic converge. Thus the optimal policy can be found by searching for the action with maximum action value for each state. After several training episodes,

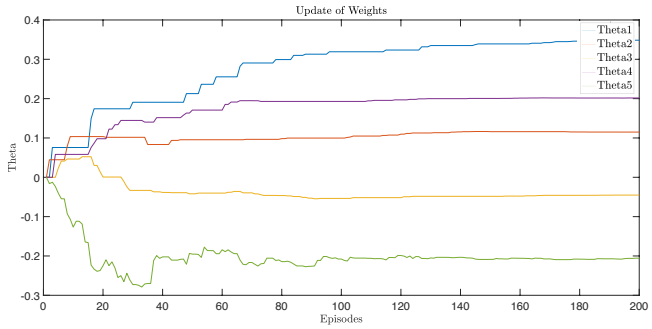

Figure 4: Update of theta elements.

the profits line still has some fluctuations due to exploration behaviours of reinforcement learning algorithm. This exploration behaviour is to ensure the algorithm can reach the optimal results and find the solution for the bidding strategy as shown in Fig. 5.

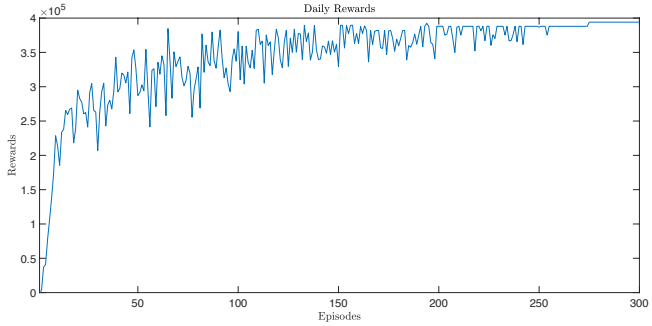

Figure 5: BESS daily rewards during the training period.

6.4. Results and Comparison

Figs. 6 - 7 show the optimal bidding strategies and bidding prices of the BESS in different time slots. In this case, regulation capacity dominates most of the day, since the compensation of the regulation services are high. Furthermore, we test bidding strategy of the BESS1 in regulation market and energy market with its rivals. To win the regulation services offer and earn high compensation profits, the bidding regulation price is trained to be less than the history clearing prices and the rivals' bids. When the regulation price are cheap, the BESS will not do much regulation mileages, so as to the BESS owner will purchase or sell the energy in the energy market to balance the energy loss and earn some revenue. During that period, the regulation bids are reduced because of the physical constraints of the charging/discharging rate.

To reveal the impact of the ageing losses and transmission losses, we magnify the loss coefficient ten times, and the simulation results are shown in Figs. 8 - 9. The impact of considering these losses on BESS can be observed by comparing Fig. 6 with Fig. 8. Due to the increase of ageing losses, the income from the regulation market is comparatively lower than that before, so that the regulation bids are decreased to extend

Table 1: Cost coefficients for simulation studies.

	ξ	C_P $(£/kW)$	C_E $(£/kWh)$	C_F $(£)$	C_a $(£/kWh)$
BESS1	15%	2300	300	2.58e5	14.6
BESS2	15%	2250	450	2.52e5	15.8
BESS3	15%	2470	360	2.49e5	16.2
BESS4	15%	2320	280	2.63e5	15.4

Table 2: Parameters of BESSs for simulation studies.

	P_{max}(MW)	E_{max}(MWh)	η_c	η_d	N_{100}^{fail}	k_p
BESS1	406	900	0.868	0.92	10,000	0.85
BESS2	207	1000	0.88	0.95	10,000	0.85
BESS3	250	625	0.86	0.88	10,000	0.85
BESS4	362	830	0.82	0.86	10,000	0.85

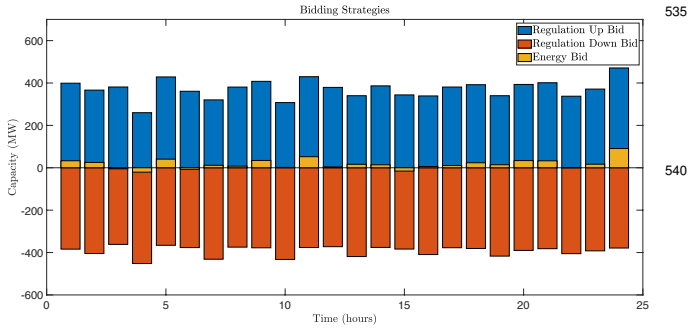

Figure 6: BESS bids.

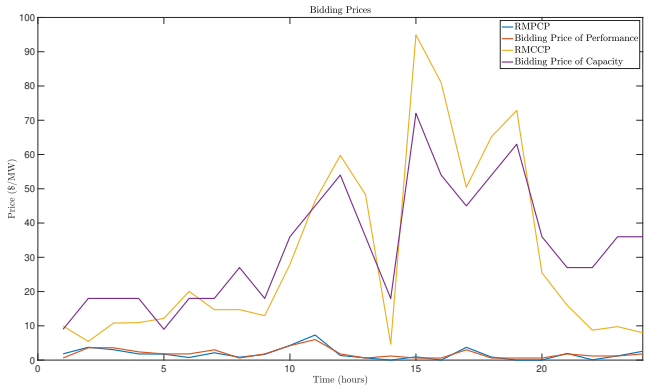

Figure 7: BESS bidding prices.

the lifetime of BESS and reduce the transmission losses. Furthermore, with the increase of transmission losses, the Energy bids are increased to balance energy losses during operations.

In Fig. 9, the bidding prices are different from the base case. Because of the high cost of losses, it is not worth operating the BESS when the prices are low. In this case, the proposed algorithm will increase the bidding prices to save the cost of regulation market. The higher bidding price will lose more frequency offers, but reduce the transmission and ageing losses.

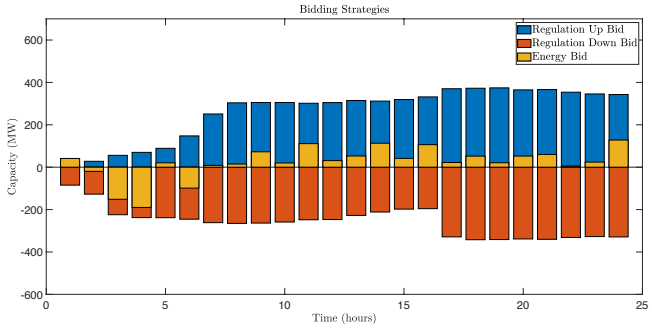

Figure 8: BESS bids with ten times losses penalty.

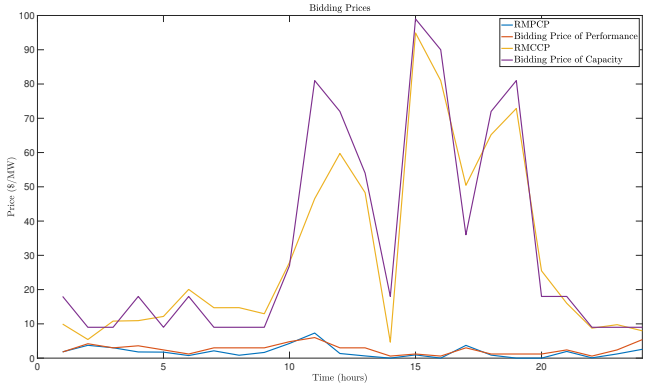

Figure 9: BESS bidding prices with ten times losses penalty.

Table. 3 summarises the profit in different markets and costs separately. It can be seen that the benefit from regulation mar-

ket is the major revenue of BESSs. For our bidding strategy in BESS 1, the BESS has to purchase electricity to balance the energy consumption and losses, so that the reward from energy market is negative. It means that the BESS would be deeply involved in regulation market to get high revenue. For other BESSs, BESS2 and BESS3 get lower rewards due to their maximum charging/discharging rate limits. The BESS4 has similar parameters as the BESS1, but earns around $70,000 less than the BESS1. This is because the proposed bidding strategy of the BESS1 can receive and learn the reward/penalty signal from the system operator, which does not require any other prior knowledge and study of its rivals. The comparison results show that the proposed model considering the ageing and transmission losses presents a more effective bidding strategy for BESS owners in a bidding environment of multiple rivals, and provides a more realistic and accurate cost-benefit result for investors as well.

Table 3: Income and cost comparison.

	BESS1	BESS2	BESS3	BESS4
Profit^{e} ($)	-17426	-1270	-8147	-14058
Profit^{cap} ($)	346673	182785	208725	284561
Profit^{perf} ($)	37327	14140	18489	23415
Cost^{total} ($)	-10676	-4803	-6419	-8218
Daily Income ($)	355898	190852	212648	285700

In addition, to further verify that the proposed algorithm can obtain the maximum profit for BESS owners, the comparison cases with different algorithms are studied and listed in Table 4. Due to the uncertainty of the bidding environment and lack of rival's information, some traditional numerical optimisation approaches, such as game theoretic [40], are not suitable for this environment. Therefore, we compare our results with other learning and stochastic optimisation approaches, Q-learning [41], State-Action-Reward-State-Action (SARSA) [42] and PSO [43]. The proposed FARL algorithm, as shown in the second column in Table 4, successfully achieves highest revenue compared with other methods. The highest incomes for each hour are highlighted in Table 4. Although Q-learning, SARSA and PSO could have higher economic performance at some time slots, our algorithm could help BESS owner get the highest profit for the majority time periods. Judging from the total income of the day, the proposed FARL has the advantage by making around 2.5%, 13% and 6.2% improvement than Q-learning, SARSA and PSO.

Table 4: Hourly income comparison.

Hour	**FARL**	Q-Learning [41]	SARSA [42]	PSO[43]
01:00	**6390**	5274	3875	5579
02:00	3707	3548	3638	**4419**
03:00	**8468**	2570	3268	2723
04:00	4526	**4633**	2855	2417
05:00	5883	6969	**8008**	4838
06:00	**11803**	8070	6901	5744
07:00	7635	16929	9934	**19091**
08:00	7622	7024	**9171**	7377
09:00	5795	4817	**13588**	6920
10:00	9162	6784	**14808**	13761
11:00	12519	**28564**	10401	13796
12:00	23876	**38742**	8322	8787
13:00	**24214**	2517	18872	16594
14:00	2149	**14809**	10586	5802
15:00	**43959**	40732	35518	42943
16:00	**42546**	21063	24420	26300
17:00	26039	27564	**32043**	23424
18:00	**34487**	18029	28299	27555
19:00	**36691**	27275	17270	32065
20:00	10878	**30830**	11964	23783
21:00	10381	10893	11153	**12171**
22:00	5289	7170	12820	**19624**
23:00	8165	8914	**10088**	6374
24:00	3714	3215	**7147**	3029
Total	**355898**	346935	314949	335116

7. Conclusion

This paper studied the optimised bidding strategy of the BESS to maximise the profits under a multi-rivals environment. We firstly proposed a bidding model for the BESS in the AGC and energy market, then solved the bidding problem with the reinforcement learning, which using function approximation to avoid aggregated states and dimension curse. Simulation results verified that the proposed method not only get the higher revenue from the AGC market, but also extends the lifetime of the BESS and reduces the losses.

References

[1] H. F. Habib, C. R. Lashway, O. A. Mohammed, A review of communication failure impacts on adaptive microgrid protection schemes and the use of energy storage as a contingency, IEEE Trans. Ind Appl. 54 (2) (2017) 1194–1207.

[2] NextEra Energy, Inc, Nextera energy annual report 2018, http://www.investor.nexteraenergy.com/~/media/Files/N/NEE-IR/reports-and-fillin\gs/annual-reports/NextEra%20Energy_Annual_Report_2018.pdf (December 2018).

[3] H. Zhao, Q. Wu, S. Hu, H. Xu, C. N. Rasmussen, Review of energy storage system for wind power integration support, Appl. energy 137 (2015) 545–553.

[4] K. Divya, J. Østergaard, Battery energy storage technology for power systems—an overview, Electr. Power Syst. Res. 79 (4) (2009) 511–520.

[5] Decision on multiple-use application issues, http://docs.cpuc.ca.gov/PublishedDocs/Publis\hed/G000/M204/K478/204478235.pdf (November 2018).

[6] Q. Shi, F. Li, Q. Hu, Z. Wang, Dynamic demand control for system frequency regulation: Concept review, algorithm comparison, and future vision, Electr. Power Syst. Res. 154 (2018) 75–87.

[7] Pjm manual 28: Operating agreement accounting, https://www.pjm.com/~/media/documents/m\anuals/m28.ashx (December 2019).

[8] B. Xu, A. Oudalov, J. Poland, A. Ulbig, G. Andersson, Bess control strategies for participating in grid frequency regulation, IFAC Proc. Vol. 47 (3) (2014) 4024–4029.

[9] M. ud din Mufti, S. A. Lone, S. J. Iqbal, M. Ahmad, M. Ismail, Supercapacitor based energy storage system for improved load frequency control, Electr. Power Syst. Res. 79 (1) (2009) 226–233.

[10] J. Tan, Y. Zhang, Coordinated control strategy of a battery energy storage system to support a wind power plant providing multi-timescale frequency ancillary services, IEEE Trans. Sustain. Energy 8 (3) (2017) 1140–1153.
[11] Y. Cheng, M. Tabrizi, M. Sahni, A. Povedano, D. Nichols, Dynamic available agc based approach for enhancing utility scale energy storage performance, IEEE Trans. Smart Grid 5 (2) (2014) 1070–1078.
[12] G. He, Q. Chen, C. Kang, Q. Xia, Optimal offering strategy for concentrating solar power plants in joint energy, reserve and regulation markets, IEEE Trans. Sustain. Energy 7 (3) (2016) 1245–1254.
[13] G. He, Q. Chen, C. Kang, P. Pinson, Q. Xia, Optimal bidding strategy of battery storage in power markets considering performance-based regulation and battery cycle life, IEEE Trans. Smart Grid 7 (5) (2015) 2359–2367.
[14] M. Marzband, M. Javadi, S. A. Pourmousavi, G. Lightbody, An advanced retail electricity market for active distribution systems and home microgrid interoperability based on game theory, Electr. Power Syst. Res. 157 (2018) 187–199.
[15] D. P. Chassin, J. C. Fuller, N. Djilali, Gridlab-d: An agent-based simulation framework for smart grids, J. Appl. Math. 2014.
[16] J. R. Vázquez-Canteli, Z. Nagy, Reinforcement learning for demand response: A review of algorithms and modeling techniques, Appl. energy 235 (2019) 1072–1089.
[17] T. Krause, E. V. Beck, R. Cherkaoui, A. Germond, G. Andersson, D. Ernst, A comparison of nash equilibria analysis and agent-based modelling for power markets, Int. J. Electr. Power Energy Syst. 28 (9) (2006) 599–607.
[18] H. Kebriaei, A. Rahimi-Kian, M. N. Ahmadabadi, Model-based and learning-based decision making in incomplete information cournot games: A state estimation approach, IEEE Trans. Syst., Man, Cybern. Syst. 45 (4) (2014) 713–718.
[19] Z. Li, Z. Ding, M. Wang, E. Oko, Model-free adaptive control for mea-based post-combustion carbon capture processes, Fuel 224 (2018) 637–643.
[20] V. Nanduri, T. K. Das, A reinforcement learning model to assess market power under auction-based energy pricing, IEEE Trans. Power Syst. 22 (1) (2007) 85–95.
[21] E. Lakić, G. Artač, A. F. Gubina, Agent-based modeling of the demand-side system reserve provision, Electr. Power Syst. Res. 124 (2015) 85–91.
[22] N. Holjevac, T. Capuder, N. Zhang, I. Kuzle, C. Kang, Corrective receding horizon scheduling of flexible distributed multi-energy microgrids, Appl. Energy 207 (2017) 176–194.
[23] R. Bellman, Dynamic programming, Science 153 (3731) (1966) 34–37.
[24] W. Tasnin, L. C. Saikia, Performance comparison of several energy storage devices in deregulated agc of a multi-area system incorporating geothermal power plant, IET Renew. Power Gener. 12 (7) (2018) 761–772.
[25] PJM Staff, Implementation and rationale for pjm's conditional neutrality regulation signals, https://www.pjm.com/~/media/committees-groups/task-forces/rmistf/postings/regula\tion-market-whitepaper.ashx (January 2017).
[26] Y. Meng, M. Liang, J. F. DeCarolis, N. Lu, Design of energy storage friendly regulation signals using empirical mode decomposition, in: 2019 IEEE Power & Energy Society General Meeting (PESGM), IEEE, 2019, pp. 1–5.
[27] Independent Market Monitor for PJM, State of the market report for pjm, volume 2: Detailed analysis, https://legalectric.org/f/2020/03/2019-som-pjm-volume2.pdf (March 2020).
[28] Y. Xu, W. Zhang, G. Hug, S. Kar, Z. Li, Cooperative control of distributed energy storage systems in a microgrid, IEEE Trans. Smart Grid 6 (1) (2014) 238–248.
[29] X. Zhou, J. L. Stein, T. Ersal, Battery state of health monitoring by estimation of the number of cyclable li-ions, Control. Eng. Pract. 66 (2017) 51–63.
[30] S. Saxena, Y. Xing, D. Kwon, M. Pecht, Accelerated degradation model for c-rate loading of lithium-ion batteries, Int. J. Electr. Power Energy Syst. 107 (2019) 438–445.
[31] B. Xu, A. Oudalov, A. Ulbig, G. Andersson, D. S. Kirschen, Modeling of lithium-ion battery degradation for cell life assessment, IEEE Trans. Smart Grid 9 (2) (2016) 1131–1140.
[32] W. Ying, Z. Zhi, A. Botterud, K. Zhang, D. Qia, Stochastic coordinated operation of wind and battery energy storage system considering battery degradation, J. Mod. Power Syst. Clean Energy 4 (4) (2016) 581–592.
[33] G. Li, J. Shi, X. Qu, Modeling methods for genco bidding strategy optimization in the liberalized electricity spot market–a state-of-the-art review, Energy 36 (8) (2011) 4686–4700.
[34] Z. Li, Z. Ding, M. Wang, Operation and bidding strategies of power plants with carbon capture, IFAC-PapersOnLine 50 (1) (2017) 3244–3249.
[35] B. Bahmani-Firouzi, R. Azizipanah-Abarghooee, Optimal sizing of battery energy storage for micro-grid operation management using a new improved bat algorithm, Int. J. Electr. Power Energy Syst. 56 (2014) 42–54.
[36] R. D. Masiello, B. Roberts, T. Sloan, Business models for deploying and operating energy storage and risk mitigation aspects, Proc. IEEE 102 (7) (2014) 1052–1064.
[37] W. Liu, P. Zhuang, H. Liang, J. Peng, Z. Huang, Distributed economic dispatch in microgrids based on cooperative reinforcement learning, IEEE Trans. Neural Netw. Learn. Syst. 29 (6) (2018) 2192–2203.
[38] M. Rashid, A. Gupta, Mathematical model for combined effect of sei formation and gas evolution in li-ion batteries, ECS Electrochem. Lett. 3 (10) (2014) A95–A98.
[39] E. Saiz-Marin, J. García-González, J. Barquin, E. Lobato, Economic assessment of the participation of wind generation in the secondary regulation market, IEEE Trans. Power Syst. 27 (2) (2012) 866–874.
[40] J. Lee, J. Guo, J. K. Choi, M. Zukerman, Distributed energy trading in microgrids: A game-theoretic model and its equilibrium analysis, IEEE Trans. Ind. Electron. 62 (6) (2015) 3524–3533.
[41] D. E. Aliabadi, M. Kaya, G. Sahin, Competition, risk and learning in electricity markets: An agent-based simulation study, Appl. energy 195 (2017) 1000–1011.
[42] Z. Li, Z. Ding, M. Wang, Optimal bidding and operation of a power plant with solvent-based carbon capture under a co2 allowance market: A solution with a reinforcement learning-based sarsa temporal-difference algorithm, Engineering 3 (2) (2017) 257–265.
[43] A. D. Yucekaya, J. Valenzuela, G. Dozier, Strategic bidding in electricity markets using particle swarm optimization, Electr. Power Syst. Res. 79 (2) (2009) 335–345.
[44] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

8. Appendix

8.1. Function Approximation

In this section, we introduce the detail of proposed reinforcement learning algorithm for BESS bidding problem.

According to the projected Bellman error J [44], the optimal updating law for the parameters θ^π can be obtained by evaluating the approximation performance. $Q^\pi(s_t, a_t)$ is simplified to Q_θ in following equations. The mean-square project Bellman error objective function can be formed as

$$J(\theta) = \|Q_\theta - \Pi T_\pi Q_\theta\|_D^2 \tag{47}$$

$$= (\Pi(T_\pi Q_\theta - Q_\theta))^T D(\Pi(T_\pi Q_\theta - Q_\theta)) \tag{48}$$

$$= (T_\pi Q_\theta - Q_\theta)^T \Pi^T D \Pi (T_\pi Q_\theta - Q_\theta) \tag{49}$$

where Π is a projection matrix which projects any action values to the linear space of approximate action values, T_π is a Bellman evaluation operator related to the Q-function Q_θ, and D is a diagonal matrix with $N \times N$ dimension, which is use to reflect the state-action pair frequency under current policy π. We have

$$\Pi = \Phi(\Phi^T D \Phi)^{-1} \Phi^T D \tag{50}$$

Further, we transfer the objective function to statistical expectation forms as

$$\begin{aligned} J(\theta) =&(T_\pi Q_\theta - Q_\theta)^T(\Phi(\Phi^T D\Phi)^{-1}\Phi^T D)^T D \\ &\Phi(\Phi^T D\Phi)^{-1}\Phi^T D(T_\pi Q_\theta - Q_\theta) && (51)\\ =&(\Phi^T D(T_\pi Q_\theta - Q_\theta))^T(\Phi^T D\Phi)^{-1}\Phi^T D(T_\pi Q_\theta - Q_\theta) && (52)\\ =&\mathbb{E}[\delta\phi]^T\mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi] && (53)\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}[\delta\phi] =& \sum_{s,a} D_{(s,a),(s,a)}\phi(s,a)\mathbb{E}[\delta_t] && (54)\\ =&\Phi^T D(T_\pi Q_\theta - Q_\theta) && (55)\end{aligned}$$

and

$$\mathbb{E}[\phi\phi^T] = \sum_{s,a} D_{(s,a),(s,a)}\phi(s,a)\phi^T(s,a) = \Phi^T D\Phi \quad (56)$$

Note that all statistical expectations are under current behaviour policy π. Also, δ is the temporal difference error, which is defined as

$$\delta_t = r_{t+1} + \gamma\hat{\phi}_t^T\theta_t - \phi_t^T\theta_t \quad (57)$$

where $\hat{\phi}_t$ is the estimated value of ϕ. In order to avoid the need for two independent samples, a modifiable parameter $w \in \mathcal{R}^n$, named quasi-stationary estimate, is introduced as follows

$$w \approx \mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi] \quad (58)$$

Then, the negative gradient of objective function can be calculated as

$$\begin{aligned} -\frac{1}{2}\nabla J_i =&\mathbb{E}[(\phi - \gamma\phi')\phi^T]\mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi] && (59)\\ =&(\mathbb{E}[\phi\phi^T] - \gamma\mathbb{E}[\phi'\phi^T])\mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi] && (60)\\ =&\mathbb{E}[\delta\phi] - \gamma\mathbb{E}[\phi'\phi^T]\mathbb{E}[\phi\phi^T]^{-1}\mathbb{E}[\delta\phi] && (61)\\ \approx&\mathbb{E}[\delta\phi] - \gamma\mathbb{E}[\phi'\phi^T]w && (62)\end{aligned}$$

Since the expectations in (59) are not know, it is generally using stochastic gradient-descent approach. To get the quicker convergence speed of the reinforcement learning algorithm, the correction term is applied to adjust the update law as follows

$$\begin{aligned}\theta_{t+1} =& \theta_t + \alpha_t(\delta_t\phi_t - \gamma w_t^T\phi_t\hat{\phi}_t) && (63)\\ w_{t+1} =& w_t + \beta_t(\delta_t - \phi_t^T w_t)\phi_t && (64)\end{aligned}$$

where $\hat{\phi}_t$ is the approximation of $\max_{a'} Q^\pi(s_{t+1}, a_{t+1})$, which can be estimated as

$$\hat{\phi}_t \approx \arg\max_{\phi(s_{t+1},a_{t+1})} \phi^T(s_{t+1}, a_{t+1})\theta_t \quad (65)$$