A Learning-Based Power Management Method for Networked Microgrids Under Incomplete Information

Qianzhi Zhang, *Student Member, IEEE*, Kaveh Dehghanpour, *Member, IEEE*, Zhaoyu Wang, *Member, IEEE*, and Qiuhua Huang, *Member, IEEE*

***Abstract*—This paper presents an approximate Reinforcement Learning (RL) methodology for bi-level power management of networked Microgrids (MG) in electric distribution systems. In practice, the cooperative agent can have limited or no knowledge of the MG asset behavior and detailed models behind the Point of Common Coupling (PCC). This makes the distribution systems unobservable and impedes conventional optimization solutions for the constrained MG power management problem. To tackle this challenge, we have proposed a bi-level RL framework in a price-based environment. At the higher level, a cooperative agent performs function approximation to predict the behavior of entities under incomplete information of MG parametric models; while at the lower level, each MG provides power-flow-constrained optimal response to price signals. The function approximation scheme is then used within an adaptive RL framework to optimize the price signal as the system load and solar generation change over time. Numerical experiments have verified that, compared to previous works in the literature, the proposed privacy-preserving learning model has better adaptability and enhanced computational speed.**

***Index Terms*—Distribution systems, networked microgrids, power management, reinforcement learning, adaptive training.**

NOMENCLATURE

Indices

i, j	Indices of bus numbers $\forall i, j \in \Omega_I$.
k	Index of line number $\forall k \in \Omega_K$.
n	Index of MG.
t	Index of episode/time instant.

Parameters

$a_f / b_f / c_f$	Coefficients of the DG quadratic cost function.
E^{Cap}	Max. capacity of ESS unit.
e_{PV}, e_D	Prediction error standard deviations.
G/B	Real/imag. parts of the bus admittance matrix.
$\hat{I}_{PV}$	Vectors of solar irradiance estimation.
I^{PV}	Real normalized solar irradiance.
$P^{Ch/Dis,M}$	Max. ESS charging/discharging limits.
P/Q^D	Active/reactive load.
$P/Q^{DG,M}$	Max. DG active/reactive power capacity.
$P^{DG,R}$	Max. DG ramp limit.
P^{PV}	PV active power output.
$P/Q^{PCC,M}$	Max. active/reactive power flow at the PCCs.
$\hat{P}_D$	Vectors of aggregate active load estimation.
P^D	Real active load.
$Q^{PV,M}$	Max. PV reactive power output limit.
S	States in Markov decision process.
L^M	Max. line flow limit.
$SOC^{M/m}$	Max./min. SOC limits.
T	Length of the moving decision window.
Δt	Time step.
α/β	Shape parameters of beta distribution.
$\eta_{Ch/Dis}$	Charging/discharging efficiency of ESS unit.
λ^F	Diesel generator fuel price.
$\lambda^{R,M/m}$	Max./min. retail price limits.
λ^W	Wholesale energy price.
θ	Vector of regression parameter.
θ^*	Vector of converged regression parameter.
θ_{Th}/V_{Th}	Threshold value.
γ	Discount factor that defines the preference.
δ	Step size that defines the rate of learning.
μ	Regularization factor.
ϕ	Forgetting factor.
ϵ	ϵ-greedy exploration factor.

Variables

a	Actions in Markov decision process.
F	Fuel consumption of DG.
SOC	SOC of the battery system.
$P^{Ch/Dis}$	Charging/discharging power of ESS unit.
P/Q^{DG}	DG active/reactive power outputs
P/Q^{ij}	Line active/reactive power flows
P/Q^{PCC}	Active/reactive power flow at the PCC.
P^W	Exchanged power with the wholesale market.
Q^{ESS}	Reactive power outputs of ESS unit.
Q^{PV}	PV inveter reactive power outputs.
$V/\Delta\theta$	Voltage magnitude and phase angle difference.

Manuscript received March 8, 2019; revised June 13, 2019 and July 25, 2019; accepted August 2, 2019. This work was supported by the U.S. Department of Energy Office of Electricity under Grant DE-OE0000839. Paper no. TSG-00346-2019. *(Corresponding author: Zhaoyu Wang.)*

AQ1

Q. Zhang, K. Dehghanpour, and Z. Wang are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: wzy@iastate.edu).

Q. Huang is with Pacific Northwest National Laboratory, Richland, WA 99354 USA (e-mail: qiuhua.huang@pnnl.gov).

AQ2

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2019.2933502

x_p/x_q	MGs power management decision vectors.
λ^R	Retail price signals at the PCCs.
$u^{Ch/Dis}$	ESS charge/discharge binary variables.

Functions

$Q_t(S, a)$	State-action value function.
$Q_t^*(S, a)$	Optimal state-action value function.
$\hat{Q}_t(S, a\|\theta)$	Parameterized approximate state-action value function.
$Q_{S\cdot a}(t\|\theta)$	Parameterized regression sub-component with state-action interaction.
$Q_S(t\|\theta)$	Parameterized regression sub-component with state values.
$Q_a(t\|\theta)$	Parameterized regression sub-component with action values.
$R(t)$	Reward function in Markov decision process.

I. INTRODUCTION

A SMART distribution system consisting of networked microgrids (MGs), with local Distributed Generators (DG), Renewable Energy Resources (RES), and Energy Storage Systems (ESS), can facilitate reliable service provision to customers in power systems [1]. Smart independent MGs are considered as a viable solution for electrification of rural areas, which are excluded from traditional electrification programs due to their remote location and financial constraints [2]. To ensure the long-term sustainability and encourage economic development in rural communities, the feasibility of cooperative business models for rural system electrification has been analyzed previously [2]–[4]. It has been shown that a non-profit cooperative can act as an intermediary agent between the rural MGs and the wholesale market. The power is exchanged between the MGs and the cooperative at a retail rate, and the revenue from electricity sales in the wholesale market is returned to MGs. The retail energy pricing program can be used to influence the MGs' behavior based on the availability of resources. Real cases of cooperative business models with rural MGs as participating members can be found in [3], [4]. The autonomous cooperative business settings in these cases have been designed to benefit rural communities.

Coordinating the real-time behavior of multiple privately-owned rural MGs in a cooperative business model is a necessary, yet challenging task [5], [6]. Due to data privacy and ownership concerns for MGs, the main difficulty in the way of obtaining a desirable coordination scheme is the limited access to real-time asset behaviors and models behind the Point of Common Coupling (PCC) with MGs, which hinders conventional model-based constrained power management solvers. This problem becomes more severe as the penetration of MGs in rural distribution systems grows. A wide range of methods have been applied in the literature with the aim of economic operation of the networked MGs, including methods such as heuristic techniques [7], [8], centralized decision models [9], [10], constrained hierarchical control architectures [11]–[13], and distributed optimization methods [14], [15].

However, the functionality of previous models [7]–[15] highly depends on the full system operator's knowledge of MG operation behind the PCC and customers' private data at node-level, including nodal demand load consumption, nodal generation capacities, nodal PV generations, sensitive cost information, asset constraints, as well as MG network topology and configuration data. Access to these information could compromise the data confidentiality and privacy of MGs and customers that participate in a cooperative business setting. Also, previous methods can be mostly categorized as "model-based", since the decision agents depend on detailed physical models of the distribution systems. One shortcoming of model-based solutions is their inability to adapt to constantly-changing system conditions when the amount of measurement data is limited.

A promising alternative to model-based optimization approaches is reinforcement learning (RL), which is a model-free data-driven technique that can be used to optimize the behavior of an agent through repeated interactions with its environment, without full system identification and no *a priori* knowledge of the system. A number of papers have given examples of how RL techniques can be applied in power systems. In [16], [17], energy consumption scheduling problems were solved for single MGs and individual residential buildings using RL algorithms. However, the above studies only focus on providing optimal solutions to power management problems for single entities instead of addressing coupled decision models for multiple interconnected entities in a cooperative setting.

In this paper, to solve the problem of decision making under incomplete information while providing decision adaptability, a bi-level cooperative framework is proposed using an RL-based method for a distribution system consisting of multiple networked privately-owned MGs: at Level I of the hierarchy, a non-profit cooperative agent maximizes the total MGs' revenue from power exchange with the wholesale market. This is done by setting the retail prices, with access only to active/reactive power measurements at the MG PCCs and *aggregate* load and solar irradiance information behind the PCCs. The cooperative agent acts as an intermediary between the MGs and the wholesale market, and returns the revenue to the MGs. At Level II of the hierarchy, each MG Control Center (MGCC) agent receives the price signal from the cooperative agent and solves the power-flow-constrained MG power management problem. The objective at this level consists of the MG operational cost and the allocated revenue from the cooperative agent. In summary, the main contributions of this paper can be listed as follows:

- The proposed power management system can handle the current limitations raised from data privacy and ownership in the cooperative setting. Considering the model-free nature of our RL-based method, the data privacy of MGs and the data confidentiality of customers are maintained. The power management problem is solved with access to only minimal and aggregated data.
- The proposed RL solver is faster than conventional optimization solvers since the learned state-action value function acts similar to a *memory* that recalls from the

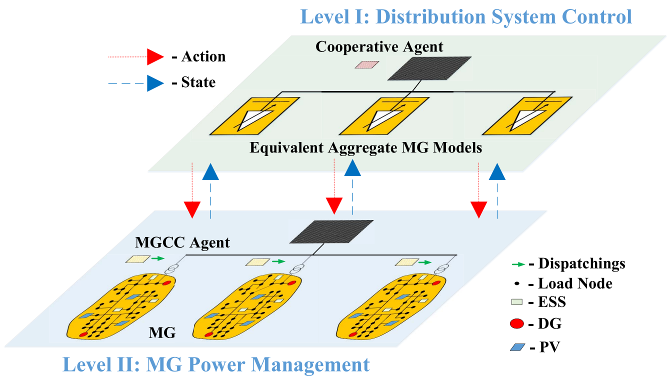


Fig. 1. The architecture of the bi-level networked MGs power management.

cooperative agent's past experiences to estimate new optimal solutions. This is done by updating the state values at each decision window and without re-solving the decision problem.

- The RL framework is trained using a regularized recursive least square methodology with a *forgetting factor*, which enables the decision model to be adaptive to changes in system parameters which are excluded from the cooperative agent's state set.

The reminder of the paper is organized as follows: Section II presents the overall decision hierarchy. Section III elaborates the proposed RL-based framework. Section IV describes the MG power management problem. Simulation results and conclusions are given in Sections V and VI, respectively.

II. Overall Decision Hierarchy

Fig. 1 gives a general overview of the proposed bi-level power management scheme for a distribution system with multiple MGs, as follows:

Level I - RL-based Distribution System Control: The cooperative agent employs an adaptive model-free RL method, developed using a regularized recursive least square function approximation methodology, to find the optimal retail price signals for the MGs based on the latest system states. This cooperative agent is *non-profit* in the sense that it does not maximize its own profit, but maximizes the social welfare for the whole system, which includes the summation of profits of all the MGs as participating members in the cooperative business model. The price signals are then transmitted to MGCC agents. The RL training process is performed by the cooperative agent through repeated interactions with the MGCC agents. At this level, each MG is modeled as an aggregate controllable load which is price-sensitive. The task of the RL algorithm is to discover the complex relationship between retail price and exchanged power with MGs at PCCs, without direct detailed knowledge of system operation behind the PCCs and only with access to estimations of the solar irradiance and aggregate fixed loads for each MG. Based on the definitions of data privacy and confidentiality in smart grid [18], this approach limits the need for access to local cost and operational constraint data of individual MGs in the first place. Hence, the proposed method maintains both the privacy of personal information and privacy of behavior for MGs. Moreover, unlike conventional centralized optimization methods, the proposed RL technique does not need customer confidential information at the node-level, such as customer load consumption, as it only uses aggregate data at the MG PCCs for optimal decision making. Furthermore, renewable and load power uncertainty are represented within the learning model state set. To facilitate adaptive conformation to changes in system parameters that are not included in cooperative agent's state set, such as fuel price, a forgetting mechanism has been integrated into the training process to assign higher importance levels to the latest observed data, compared to previous observations.

Level II - MG Power Management: At the second level, the MGCC agents receive the price signal for a look-ahead moving decision window. Based on the received price signals, each MGCC agent solves a constrained Mixed Integer Nonlinear Programming (MINP) to dispatch their local generation/storage assets to maximize their revenue (or equivalently minimize their cost) in the price-based environment, subject to full AC power flow constraints. Each MG's total revenue includes the cost of operation and the allocated revenue received from the cooperative agent. Based on the solution to this problem, each individual MGCC agent determines the exchanged active and reactive power with the distribution system at PCC.

Note that the RL-based reward maximization problem at Level I is subject to the power-flow-constrained response of MGs at Level II. Since the MGs are sensitive to electricity price, the reward value cannot be maximized by setting the price to its highest value. This will lead to the maximum DG generation, which will result in a decline in the cooperative agent's revenue. Hence, optimal price is reached based on a tradeoff between MGs' over-generation (when price is too high) and over-consumption (when price is too low). Also, note that the response of MGs itself is explicitly constrained by network power flow constraints.

III. Level I: Adaptive RL-Based Distribution System Control

At the first level of the hierarchy, a non-profit cooperative agent is in charge of setting the retail price of electricity at different times to maximize the revenue from power exchange with wholesale market, which will be allocated between MGs. This problem is formulated and solved over a moving decision window of length T. The difficulty in solving this problem is that the cooperative agent has incomplete knowledge of MGs' asset control and management data. To solve this problem, an RL approach is adopted, in which the decision making cooperative agent observes the response of its environment, consisting of networked MGs, to its actions at different states. Based on the received reward/cost signals from its environment and without explicit modeling, the cooperative agent searches for actions that optimize its expected accumulated received rewards at different system states.

A. Proposed RL-Based Method Structure

A RL framework consists of a Markov decision process including a set of states ($\boldsymbol{S} \in \mathcal{S}$), a set of actions ($\boldsymbol{a} \in \mathcal{A}$), a reward function ($\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$), and a state-action value function corresponding to each state-action pair ($Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$). These components are defined for the problem at hand, as follows.

1) State Set Definition: In this paper, the system state, which is denoted by $\boldsymbol{S}(t) = (\boldsymbol{S_1}(t), \ldots, \boldsymbol{S_N}(t))^\top$ at time t, is a concatenation of MGs' local state vectors ($\boldsymbol{S_n}(t)$ for the n^{th} MG) defined as:

$$\boldsymbol{S_n}(t) = \left\{\hat{I}_{PV}(t, n), \hat{P}_D(t, n)\right\} \tag{1}$$

where, $\hat{I}_{PV}(t, n)$, $\hat{P}_D(t, n)$ are the vectors of solar irradiance estimation, and aggregate active load power estimation for the n^{th} MG at time t, respectively. Hence, to define the global state, the cooperative agent needs to estimate or predict the uncertain aggregate solar irradiance and load at the PCC for each MG. To represent the uncertainty of the prediction process, prediction error values are considered to the actual underlying solar irradiance and load values, as shown below:

$$\hat{I}_{PV}(t, n) \sim Beta(\alpha, \beta) \tag{2a}$$

$$\alpha = \frac{\beta\left(\sum_i I^{PV}_{i,t,n}\right)}{\left(1 - \sum_i I^{PV}_{i,t,n}\right)} \tag{2b}$$

$$\beta = \left(1 - \sum_i I^{PV}_{i,t,n}\right)\left(\frac{\sum_i I^{PV}_{i,t,n}\left(1 + \sum_i I^{PV}_{i,t,n}\right)}{e^2_{PV}} - 1\right) \tag{2c}$$

$$\hat{P}_D(t, n) \sim \mathcal{N}\left(\sum_i P^D_{i,t,n}, e^2_D(t)\right) \tag{2d}$$

where, $\sum_i I^{PV}_{i,t,n}$ and $\sum_i P^D_{i,t,n}$ are the real aggregate normalized solar irradiance and load over the decision window, and e_{PV} and e_D are the beta and Gaussian estimation error standard deviations. The values of parameters of beta and Gaussian distributions are adopted from the [19]–[21].

2) Action Set Definition: Given the definition of model states, the global action vector is similarly defined by the retail price signals at the PCCs with MGs, denoted as $\lambda^R_{t,n}$ for the n^{th} MG, $\boldsymbol{a}(t) = (\lambda^R_{t,1}, \ldots, \lambda^R_{t,N})^\top$.

3) Reward Function Definition: The reward function at time t represents the discounted accumulated revenue of the cooperative agent over the moving decision window with length T:

$$R(t) = \sum_{t'=0}^{T-1} \gamma^{t'} \left(\lambda^W_{t+t'} P^W_{t+t'} - \sum_{n=1}^{N} \lambda^R_{t+t',n} P^{PCC}_{t+t',n}\right) \tag{3}$$

where, γ is a discount factor ($0 \leq \gamma \leq 1$) that defines the cooperative agent's preference for the immediate reward, defined as the revenue at time t, $\pi(t) = \lambda^W_t P^W_t - \sum_{n=1}^{N} \lambda^R_{t,n} P^{PCC}_{t,n}$. Also, λ^W_t denotes the wholesale energy price, P^W_t is the exchanged power with the wholesale market, where $P^W_t \leq 0$ represents power import from the wholesale market. $P^{PCC}_{t,n}$ is the active power transfer between grid and the n^{th} MG through the PCC, where $P^{PCC}_{t,n} \geq 0$ implies export from MGs to grid. The extreme case of $\gamma = 0$ represents a myopic cooperative agent, which favors only the immediate economic rewards and assigns zero weights to future expected rewards. However, as the discount factor increases the cooperative agent starts to include future expected rewards into its optimal decision problem. Hence, when the discount factor reaches $\gamma = 1$ the cooperative agent assigns equal weights to all the expected reward values for all the time instants in the decision window.

4) State-Action Value Function Parameterization: To optimize the cooperative agent's action, an auxiliary state-action value function is formed, denoted as $Q(S, a)$, which can be thought of as a replacement for the explicit system model. The state-action value function determines the long-term accumulated expected reward given the current state and action vectors:

$$Q_t(\boldsymbol{S}, \boldsymbol{a}) = E\left\{\sum_{t'=0}^{T-1} \gamma^{t'} \pi\left(t + t'\right) | \boldsymbol{S}(t) = \boldsymbol{S}, \boldsymbol{a}(t) = \boldsymbol{a}\right\} \tag{4}$$

where, $Q_t(\boldsymbol{S}, \boldsymbol{a})$ is the expected accumulated reward if the initial starting state is $\boldsymbol{S}(t)$, while the selected initial action is $\boldsymbol{a}(t)$, and the latest optimal policy is followed for every other time-step in the future. The expectation operator $E\{\}$ is calculated with respect to the future expected action-states, which in this case are in turn functions of the solar-load uncertain powers.

The goal of RL is to learn an optimal state-action value function, $Q^*_t(\boldsymbol{S}, \boldsymbol{a})$, that satisfies the Bellman optimality equation [22], as follows:

$$Q^*_t(\boldsymbol{S}, \boldsymbol{a}) = E\left\{\pi(t + 1) + \gamma \cdot \max_{\boldsymbol{a}'} Q^*_t\left(\boldsymbol{S}(t + 1), \boldsymbol{a}'\right)\right\} \tag{5}$$

Since solving (5) directly is not possible, RL provides a framework to obtain the optimal state-action value function which satisfies (5) using an iterative episodic learning environment. To implement this framework for the cooperative agent interacting with multiple MGs, the state-action value function is parameterized employing a multivariate polynomial regression approximation technique [22]–[24], defined by $\hat{Q}_t$, which consists of three multivariate polynomial elements with maximum degree 2:

$$Q_t(\boldsymbol{S}, \boldsymbol{a}) \approx \hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta}) = Q_{\boldsymbol{S}\cdot\boldsymbol{a}}(t|\boldsymbol{\theta}) + Q_{\boldsymbol{S}}(t|\boldsymbol{\theta}) + Q_{\boldsymbol{a}}(t|\boldsymbol{\theta}) \tag{6}$$

Given the regression parameter vector $\boldsymbol{\theta}$, $Q_{\boldsymbol{S}\cdot\boldsymbol{a}}$, $Q_{\boldsymbol{S}}$, and $Q_{\boldsymbol{a}}$ are the parameterized sub-components that quantify the impacts of state-action interaction $Q_{\boldsymbol{S}\cdot\boldsymbol{a}}(t|\boldsymbol{\theta})$, state values $Q_{\boldsymbol{S}}(t|\boldsymbol{\theta})$, and action values $Q_{\boldsymbol{a}}(t|\boldsymbol{\theta})$, respectively. These regression sub-components in multivariate polynomial regression model are defined as follows:

$$Q_{\boldsymbol{S}\cdot\boldsymbol{a}}(t|\boldsymbol{\theta}) = \sum_{n=1}^{N} \theta^1_{t,n} \lambda^R_{t,n} \hat{I}_{PV}(t, n) + \sum_{n=1}^{N} \theta^2_{t,n} \lambda^R_{t,n} \hat{P}_D(t, n) \tag{7}$$

$$Q_{\boldsymbol{S}}(t|\boldsymbol{\theta}) = \sum_{n=1}^{N} \theta^3_{t,n} \hat{I}_{PV}(t, n) + \sum_{n=1}^{N} \theta^4_{t,n} \hat{P}_D(t, n) \tag{8}$$

$$Q_{\boldsymbol{a}}(t|\boldsymbol{\theta}) = \sum_{n=1}^{N} \theta^5_{t,n} \lambda^R_{t,n} + \theta^6 \tag{9}$$

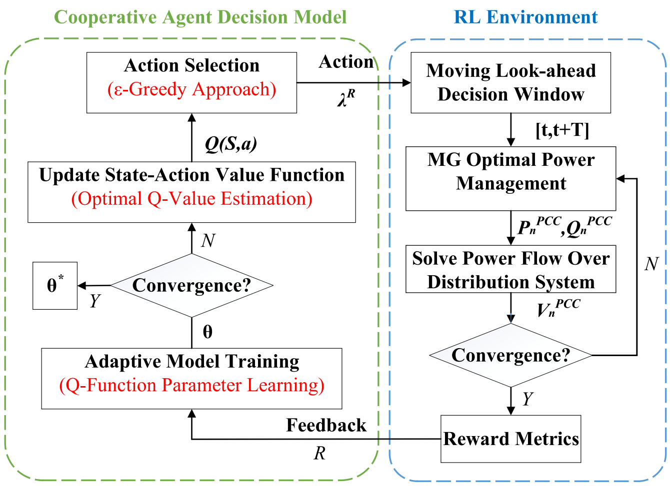


Fig. 2. Proposed RL-based framework.

where, $\boldsymbol{\theta} = \{\theta^k_{t,n}, \theta^k\}$ constitute the parameters of the approximate state-action value function that have to be learned by the cooperative agent through repeated interaction with the MGs.

Together these three components form a bilinear regression model to parametrize the state-action value function (i.e., the regression model is linear with respect to each of its arguments.) The reason for selecting a bilinear regression model is the structure of the reward function (3), which also follows a bilinear relationship between the price signal and the aggregate power measured at MG PCCs and the substation. Furthermore, the state-action value parameterization shown in (7)-(9) offers two critical advantages compared to other types of function approximators: 1) Using a bilinear regression model will simplify optimal action selection procedure considerably, as will be shown in Section III-B. For instance, if an artificial neural network is used, optimal action selection becomes intractable. However, using the proposed bilinear regression model, optimal action selection reduces to linear programming, which can be solved easily. 2) A basic challenge in choosing the form of a function approximator is the trade-off between over-parametrization and estimation accuracy. For example, as we increase the degree of the multivariate polynomial approximator the value estimation accuracy for new state-action pairs would also improve; however, at some point the function approximator becomes over-parameterized and will start overfitting to the available data, at which point the performance declines. We observed that by limiting the degree of the multivariate polynomial degree to 2, the best estimation accuracy can be achieved while maintaining a safe margin to avoid overfitting under various practical case studies.

B. Adaptive RL-Based Method Training

To achieve this task we have adopted an adaptive episodic learning mechanism, which is shown in Fig. 2. Each episode in the learning process corresponds to an online decision instant. Hence, as the decision window rolls along time new episodes are perceived by the cooperative agent. The learning process has the following steps.

Step 1 (Initialization): The time index is initialized as $t = t_0$, representing the first episode. The parameters of the state-action value function are initialized, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}(t_0)$. The initial state of the system, corresponding to solar irradiance and aggregate load of all the MGs for the decision window $[t_0, t_0 + T]$ is predicted, $\boldsymbol{S}(t_0), \ldots, \boldsymbol{S}(t_0 + T)$. Note that these predicted states, while representing system uncertainty, are updated continuously as the decision window rolls along time.

Step 2 (ϵ-greedy Action Selection): Based on the latest state-action value function defined by parameter $\boldsymbol{\theta}$, the optimal actions are estimated for the decision window $[t, t + T]$ to maximize the cooperative agent's accumulated reward, as follows:

$$\begin{aligned} \boldsymbol{a_{opt}}(t') = \arg\max_{\boldsymbol{a'}} \;\; & Q_{t'}\left(\boldsymbol{S}(t'), \boldsymbol{a'}\right) \\ s.t. \;\; & \boldsymbol{a'} = \left(\lambda^R_{t',1}, \ldots, \lambda^R_{t',N}\right)^\top \\ & \lambda^{R,m} \le \lambda^R_{t',i} \le \lambda^{R,M}, \forall\, i = \{1, \ldots, N\} \\ & \forall t' = \{t, \ldots, t + T\} \end{aligned} \tag{10}$$

where, $\boldsymbol{\rho_\lambda} = [\lambda^{R,m}, \lambda^{R,M}]$ defines the minimum/maximum range of action for retail price. Note that given the parameterization for $Q_t(\boldsymbol{S}, \boldsymbol{a})$ in (7)-(9), (10) is basically a set of linear programs, which can be solved efficiently using off the shelf solvers. A critical aspect of (10) is that the obtained optimal action, $\boldsymbol{a_{opt}}(t)$, is calculated with respect to the latest state-action value function, which could be far from being accurate in the early stages of training. Hence, to reduce the risk of sub-optimality and to strike a balance between exploration and exploitation of decision space, an ϵ-greedy action selection method [22] is adopted, with $0 \le \epsilon \ll 1$, to select the cooperative agent's action at time t:

$$\boldsymbol{a}(t) = \begin{cases} \boldsymbol{a_{opt}}(t) & \text{if } r \ge \epsilon \\ \lambda^R_{t,i} \sim U\{\boldsymbol{\rho_\lambda}\}\ \forall i & \text{if } r < \epsilon \end{cases} \tag{11}$$

where, r is a random number selected uniformly, $r \sim U\{[0, 1]\}$, with U$\{\mathbf{A}\}$ representing uniform probability distribution over the set $\boldsymbol{A}$. The randomization (11) promotes continuous exploration of action space to improve the outcome of the learning process. Upon obtaining the action vector $\boldsymbol{a}(t)$, retail price signals are sent to each MGCC agent.

Step 3 (Networked MG Power Management): Based on the received price signals, $\lambda^R_{t',n}, \forall n, t' = \{t, \ldots, t + T\}$, each MGCC agent solves its optimal power management problem (Section IV). Based on the solutions at this stage, the aggregate power injection/withdrawal to/from the grid are obtained at the PCCs with the MGs, denoted as $P^{PCC}_{t',n}$ and $Q^{PCC}_{t',n}$, $\forall n, t' = \{t, \ldots, t + T\}$.

Step 4 (Accumulated Reward Calculation): Based on the outcomes of the MG power managements, the net power exchange with the wholesale market, P^W_t, is determined and used to calculate the discounted accumulated revenue for the decision window $[t, t + T]$, using (3).

Step 5 (Adaptive Model Training): Using the observed reward signal, the regression models defined in (7)-(9) are updated, based on a gradient descent approach to modify the parameters in the direction of improving the generalization capacity of the state-action value function [22]:

$$\boldsymbol{\theta}(t+1) \leftarrow \boldsymbol{\theta}(t) + \delta\Big\{R(t) - \hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta})\Big\}\nabla_{\boldsymbol{\theta}}\hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta}) \quad (12)$$

where, δ is the step size that defines the rate of learning. Note that ideally we require $\hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta}) = R(t)$, which implies that the approximate state-action value function is able to accurately predict the accumulated reward. Accordingly, (12) is devised to reduce this prediction error over time. To implement (12), two points have to be taken under consideration: 1) since data acquisition and the training process both depend on cooperative agent action selection, approximate RL algorithms are known to be prone to overfitting and over-estimation of the values of state-action pairs [25]. Hence, a regularization mechanism has to be adopted to reduce the risk of overfitting, 2) the distribution system parameters are subject to change over time. These time-varying parameters, such as price of fuel, are not directly captured in the Markov decision process's state definition. This makes the learned model susceptible to failure in case considerable changes occur in the values of these parameters. Hence, the training process needs to be *adaptive* to enable cooperative agent to quickly conform to new system conditions. To implement (12) while considering the above-mentioned points, a regularized recursive least squares algorithm with exponential forgetting is designed [26]. The regression parameters are updated recursively, as follows:

$$\boldsymbol{\theta}(t+1) \leftarrow \boldsymbol{\theta}(t) + \boldsymbol{\Delta}(t)\boldsymbol{x}(t)\Big\{R(t) - \hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta})\Big\} \quad (13)$$

$$\boldsymbol{\Delta}(t+1) \leftarrow \hat{\boldsymbol{\Delta}}(t+1)\Big(I + \mu\hat{\boldsymbol{\Delta}}(t+1)\Big)^{-1} \quad (14)$$

$$\hat{\boldsymbol{\Delta}}(t+1) \leftarrow \frac{1}{1-\phi}\left(\boldsymbol{\Delta}(t) - \frac{\boldsymbol{\Delta}(t)\boldsymbol{x}(t)\boldsymbol{x}(t)^{\top}\boldsymbol{\Delta}(t)}{1 + \boldsymbol{x}(t)^{\top}\boldsymbol{\Delta}(t)\boldsymbol{x}(t)}\right) \quad (15)$$

where, $\boldsymbol{x}(t) = (\boldsymbol{S}(t), \boldsymbol{a}(t))^{\top}$ represents the latest cooperative agent's observation, $\boldsymbol{\Delta}$ is an auxiliary matrix mimicking the regression pseudo-inverse matrix, μ is the regularization factor which is used for re-scaling the model covariance, and $0 \leq \phi < 1$ is the forgetting factor. The regularization factor acts as a weight for penalizing the Euclidean norm of parameter vector (i.e., $||\boldsymbol{\theta}||_2$) in a ridge regression setting to prevent overfitting. The forgetting factor enables the cooperative agent to "forget" its earlier experiences in favor of the newer observations by assigning lower weights to the previously learned parameters. Hence, the forgetting factor introduces an exponential extenuation of data history over time.

Step 6 (State Transition): The decision window is moved forward to the new episode, $t \leftarrow t+1$. The new system state for the decision window, $[t, t+T]$ is predicted and denoted as $\{\boldsymbol{S}(t), \ldots, \boldsymbol{S}(t+T)\}$.

IV. Level II: MGCC Agent Power Management

At Level II, each MG receives the price signals from the cooperative agent to solve the constrained optimal power management problem within a moving decision window individually, as shown in the paper Appendix, (16)-(40). Each MG is comprised of local DGs, ESS, solar Photo-Voltaic (PV) panels and a number of loads. Hence, to account for the impacts of MGs on each other, the MG-level optimal power flow solver is based on an interactive non-linear programming algorithm. The steps of the interactive power flow solution are as follows:

Step I (Receive Input Signals From Level I): The MGs receive the retail price signals at the PCCs, $\lambda^R_{t,n}$, from the cooperative agent.

Step II (Solve Individual MG Optimal Power Management Problem): Given $\lambda^R_{t,n}$ and the estimated voltage at PCC, the power management problem (16)-(40) is solved independently by each MGCC, and the exchanged active and reactive powers at the PCCs are obtained for each MG.

Step III (Solve Power Flow Problem Over Distribution System): Treating MGs as fixed PQ loads in the external distribution system, power flow is solved over the network connecting the MGs. The total substation exchanged power, P^W_t, and voltage values at PCCs, $V^{PCC}_{t,n}$, are updated based on the power flow solution.

Step IV (Check Convergence): Go back to Step III to update PQ values corresponding to each MG, until the changes in voltage values at MG PCCs are smaller than a threshold value V_{Th}.

To summarize, the pseudo-code of the proposed bi-level RL-based framework has been shown in Algorithm 1.

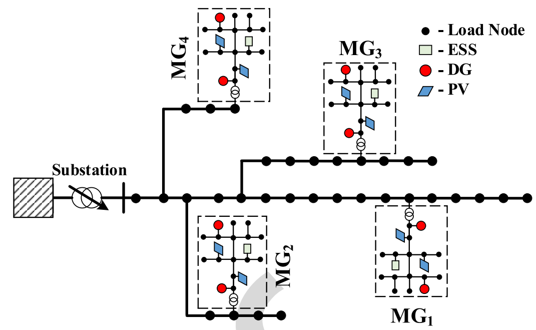


Fig. 3. Test system under study.

TABLE I
RL-Based Method Parameters

Parameters	γ	δ	μ	ϕ	ϵ
Values	0.99	0.01	1×10^{-5}	0.01	0.1

V. Numerical Results

The proposed method is tested on a modified medium voltage 33-bus distribution network [27], which has been widely used for studies pertaining to distribution system [28]. The case study consists of four MGs as shown in Fig. 3. Each MG is modeled as a modified IEEE 13-bus network at a low voltage level [29]. Hence, the system has a total number of 85 nodes. To represent a realistic model, we simulated an unbalanced system, where the loads and generators are almost uniformly distributed across phases. Note that the proposed model-free power management technique applies to both balanced and unbalanced systems. Table I presents all setting parameters for the proposed RL-based method in this paper.

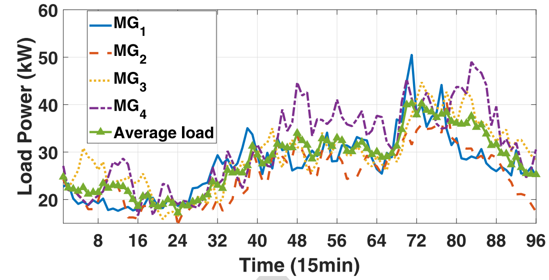
Algorithm 1 Bi-Level RL-Based Power Management Method

1: Select $T, \gamma, \delta, \mu, \phi, \epsilon, \boldsymbol{\theta}(t_0)$
2: **procedure** LEVEL I: RL ACTION SELECTION($\boldsymbol{\theta}$)
3: $\quad t \leftarrow 1$
4: $\quad \boldsymbol{S} \leftarrow [\boldsymbol{S}(t), \dots, \boldsymbol{S}(t+T)]$
5: $\quad Q_t(\boldsymbol{S}, \boldsymbol{a}) \leftarrow \hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta})$
6: $\quad \boldsymbol{a_{opt}}(t) \leftarrow$ Solve linear program (10)
7: $\quad \lambda^R_{t,i} \sim U\{\boldsymbol{\rho_\lambda}\}$
8: $\quad r \sim U\{[0, 1]\}$
9: $\quad$ **if** $r \geq \epsilon$ **then**
10: $\quad\quad \boldsymbol{a}(t) \leftarrow \boldsymbol{a_{opt}}(t)$
11: $\quad$ **else**
12: $\quad\quad \boldsymbol{a}(t) \leftarrow \lambda^R_{t,i}$
13: $\quad$ **end if**
14: **end procedure**
15: **procedure** LEVEL II: MGCC AGENT POWER MANAGEMENT($\boldsymbol{a}$)
16: $\quad k \leftarrow 1$
17: $\quad \lambda^R \leftarrow \boldsymbol{a}(t)$, $V_n(k) \leftarrow V^{PCC}_{t,n}$
18: $\quad P^{PCC}_{t,n}, Q^{PCC}_{t,n} \leftarrow$ Solve (16)-(40) $\forall n$ with $V_n(k)$
19: $\quad V_n(k) \leftarrow$ Solve power flow with $\{P^{PCC}_{t,n}, Q^{PCC}_{t,n}\}$
20: $\quad$ **if** $\Delta|V_n| \geq V_{Th}$ **then**
21: $\quad\quad k \leftarrow k + 1$
22: $\quad\quad$ Go back to Step 18
23: $\quad$ **else**
24: $\quad\quad$ Go to Step 27
25: $\quad$ **end if**
26: **end procedure**
27: **procedure** LEVEL I: RL UPDATE STATE-ACTION VALUE FUNCTION($P^{PCC}, P^W, \boldsymbol{S}, \boldsymbol{a}, \boldsymbol{\theta}$)
28: $\quad R(t) \leftarrow \sum_{t'=0}^{T-1} \gamma^{t'} (\lambda^W_{t+t'} P^W_{t+t'} - \sum_{n=1}^{N} \lambda^R_{t+t',n} P^{PCC}_{t+t',n})$
29: $\quad \hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta}) \leftarrow Q_{\boldsymbol{S}\cdot\boldsymbol{a}}(t|\boldsymbol{\theta}) + Q_{\boldsymbol{S}}(t|\boldsymbol{\theta}) + Q_{\boldsymbol{a}}(t|\boldsymbol{\theta})$
30: $\quad \boldsymbol{\theta}(t+1) \leftarrow \boldsymbol{\theta}(t) + \delta\{R(t) - \hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta})\}\nabla_{\boldsymbol{\theta}} \hat{Q}_t(\boldsymbol{S}, \boldsymbol{a}|\boldsymbol{\theta})$
31: $\quad$ **if** $||\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t)|| \geq \theta_{Th}$ **then**
32: $\quad\quad t \leftarrow t + 1$
33: $\quad\quad$ Go back to Step 4
34: $\quad$ **else**
35: $\quad\quad \theta^* \leftarrow \theta(t+1)$
36: $\quad\quad$ Output θ^*
37: $\quad$ **end if**
38: **end procedure**

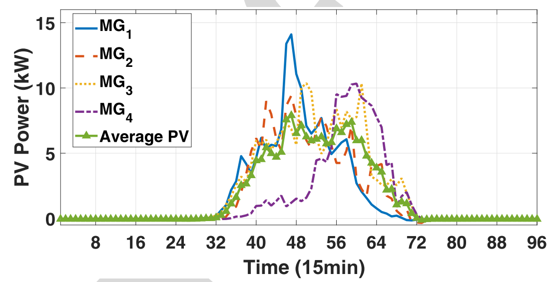
A. System Operation Outcomes

The *aggregate* active load profiles of all the MGs and the average load are presented in Fig. 4(a). The *aggregate* solar active generations in each MGs have been shown in Fig. 4(b). Both load demands and PV generations data with 15 minutes time resolution are obtained from smart meters to provide realistic numerical experiments. The wholesale market prices used in the numerical case study have been shown in Fig. 4(c), which are adopted from the historical wholesale electricity market data from U.S. Energy Information Administration [30].

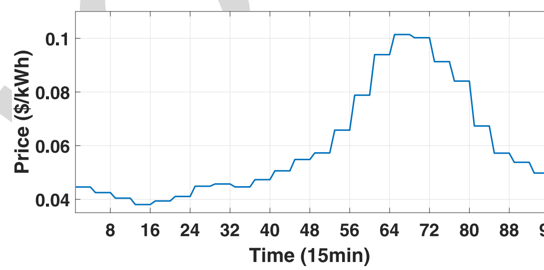
The retail price signals for the MGs, which are the optimal actions from Level I of the proposed RL-based model, are presented in Fig. 5. Power exchange between MGs and the main grid under optimal price actions, which are the responses of each MG to the actions, are shown in Fig. 6. These figures show the correlation between MGs' behavior and the retail price signal. This demonstrates the mutual impacts of the two levels of the decision model. As the wholesale price increases, the cooperative agent increases the retail prices to encourage the MGs to produce more power to reduce the costs of power purchase from the wholesale market. It can be observed that, most of the time, the cooperative agent exports power to the heavily loaded MGs to maintain power balance in the system. The reason for this is that MGs cannot provide their local demand consumption by their own local generation and have to purchase power from the cooperative service provider. The overall operational costs of MGs have been compared with and without a cooperative agent as an intermediary between the wholesale market and MGs. As can be seen from Fig. 7, the total operational costs of each MG are reduced due to the returned revenue from the cooperative service provider. Therefore, as an intermediary between the MGs and the wholesale market, the cooperative agent can help MGs to reduce their overall operational cost. Hence, it is in the interest of the MGs to participate in the wholesale market through the non-profit cooperative agent.



(a) Aggregate active load profile of the MGs



(b) Aggregate PV power of the MGs



(c) Wholesale market price

Fig. 4. Input data for the case study.

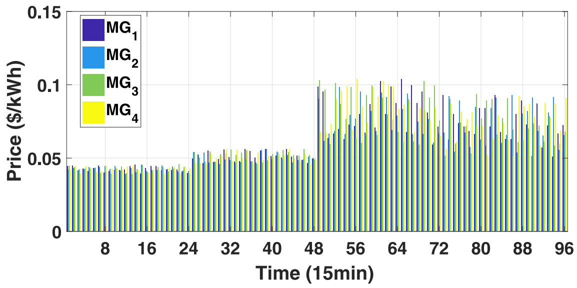


Fig. 5. Optimal retail price signals (Level I actions).

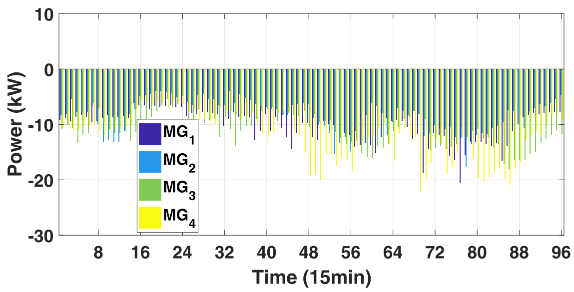


Fig. 6. Optimal power transfer through PCC of MGs (Level II responses to optimal actions).

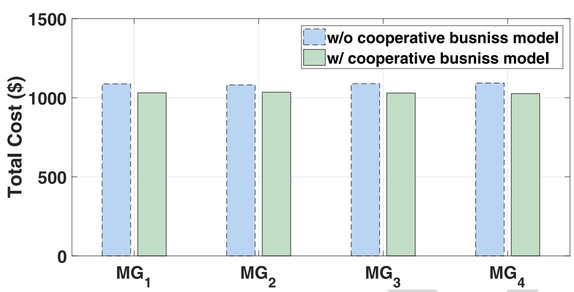


Fig. 7. Comparison of total operational cost of MGs.

B. Benefits of RL-Based Method

A numerical comparison between a centralized off the shelf solver [31] versus the proposed method for the multiple MGs power management problem is shown in Table II. In this table, the total social welfare is defined as the summation of the cooperative agent's accumulated reward and the operational cost of all the MGs. Ideally both of the solvers should output the global optimal solution to the problem. As can be seen, the difference between the solutions obtained by the centralized solver with complete system information, and the proposed RL method under incomplete information is less than 0.5% of the total achieved welfare. Note that while the initial RL training stage can be time-consuming, the decision time is much smaller than that of a centralized optimization method, upon convergence. This is due to the fact that the proposed RL-based method is able to receive continual updates over time, which enables the decision framework to reach a solution in real-time without the need to solve a large-scale optimization problem at each time instant.

To further demonstrate this, we have performed numerical experiments in which the trained state-action value functions of three different decision windows have been used for a new decision window without re-training. In Fig. 8, optimal power transfers are compared for four scenarios representing four distinct decision windows: in each scenario the RL training is performed for one of the decision windows from random initial conditions, while the updated aggregate MG solar generation and load demand from that decision window are simply inserted into the learned state-action value functions obtained from the other three decision windows. Then, the optimal actions are calculated for each decision window. As can be seen, for all scenarios the optimal solutions are close to each other and almost identical. This shows that the state-action value function learned from other decision windows can be used reliably in new situations using updated state information. Hence, the RL model does not necessarily need to be trained from scratch, and the latest learned function approximator can be simply used to update the cooperative agent's decisions. In practice, however, the re-training process has to be performed with a user-defined frequency depending on the rate of change of system parameters.

TABLE II
COMPARISON WITH A CENTRALIZED OPTIMIZATION METHOD

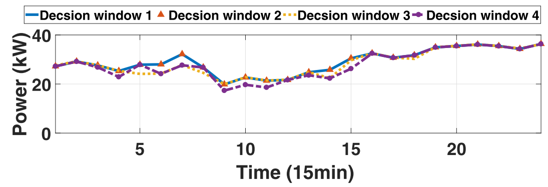
	RL-based method	Centralized Opt.
Social welfare ($)	4232.264	4212.372
Computational time (s)	9.64	116.35
MG privacy maintenance	Yes	No

Therefore, the RL-based method has two fundamental advantages over centralized optimization method: 1) RL is model-free; hence, unlike centralized optimization approaches, it does not require detailed private knowledge of MG systems to reach the optimal solution. 2) RL is much faster compared to centralized solvers since the learned state-action value function, which acts similar to a *memory*, is able to leverage the cooperative agents past experiences to obtain new optimal solutions by generalizing to new unseen states.

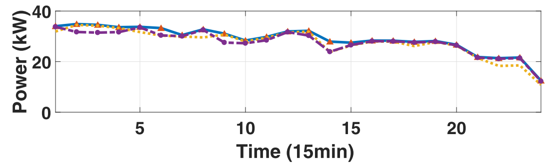
C. Adaptive RL Results

To verify the functionality of the RL framework, the estimated reward obtained from the multiple linear regression is compared with the actual reward at each episode, as shown in Fig. 9. As can be seen, at the earlier stages of the learning process, the difference between the estimated reward and the real reward is relatively high. However, as the number of episodes increases, this difference drops to within an acceptable range. The results imply that the cooperative agent is able to accurately estimate the response of MGs to control actions. Hence, using the proposed RL approach the cooperative agent is able to track the behavior of MGs and maximize the reward through continuous interactions.

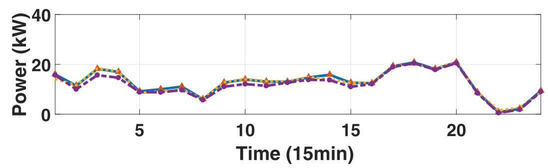
To test the adaptability of the learning framework against changes in parameters that have not been included in the definition of state set and are not directly observed by the cooperative agent, a numerical scenario is devised. At a point in time (episode $t = 250\ h$), the DG fuel price is doubled. The reward estimation mean absolute percentage error (MAPE)



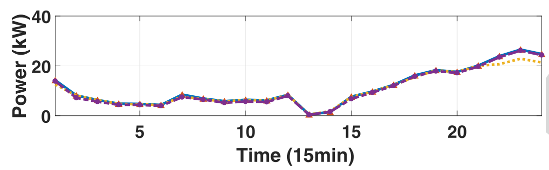
(a) Optimal action for decision window 1, using the trained models of decision windows 2, 3, and 4 for comparison



(b) Optimal action for decision window 2, using the trained models of decision windows 1, 3, and 4 for comparison



(c) Optimal action for decision window 3, using the trained models of decision windows 1, 2, and 4 for comparison



(d) Optimal action for decision window 4, using the trained models of decision windows 1, 2, and 3 for comparison

Fig. 8. Verifying the accuracy of previously-learned models under new state scenarios from different decision windows (memory effect).

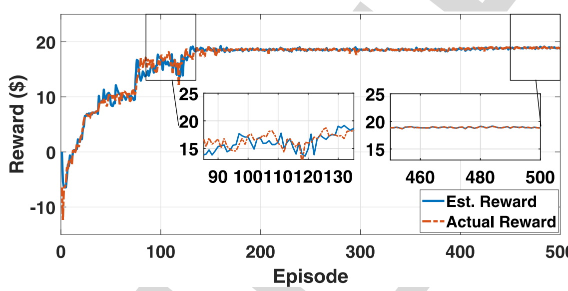
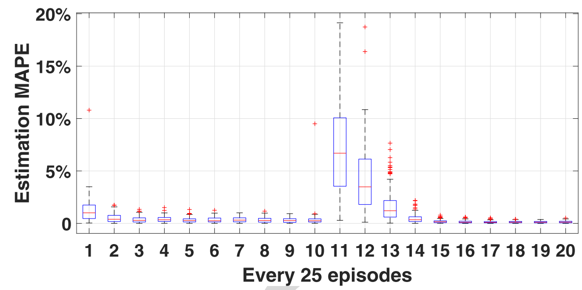
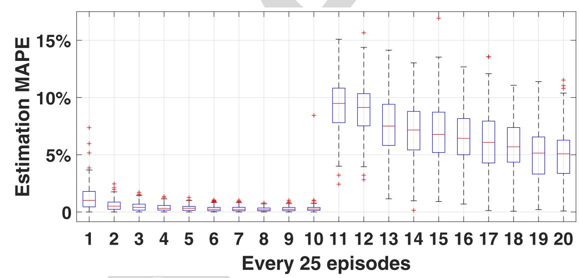


Fig. 9. Performance of the proposed reward function approximation.

with forgetting factor is shown in Fig. 10(a). As can be seen, upon the occurrence of the sudden change in fuel price, the learning MAPE temporarily jumps to a very high value since the cooperative agent is now facing a new unknown environment, as the price of fuel is not included within the cooperative agent's Markov decision process. However, as the learning process with forgetting proceeds, the MAPE drops to within acceptable range once more. The cooperative agent can still track the actual underlying reward signal as the number of episodes increases with the sudden parameter changes. The reward estimation MAPE without forgetting factor is shown



(a) Estimation MAPE with forgetting factor (Highly adaptive)



(b) Estimation MAPE without forgetting factor (Slow adaptation)

Fig. 10. Adaptability of the proposed RL-based method.

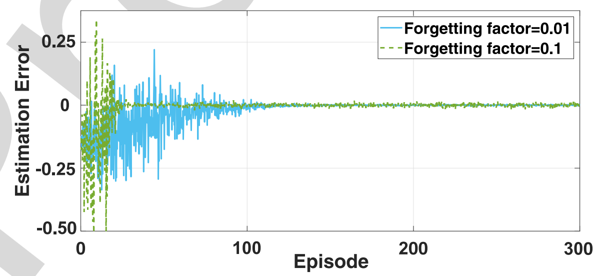


Fig. 11. Impact of forgetting factor on learning convergence.

in Fig. 10(b). As can be seen, compared to the proposed adaptive RL-based method with forgetting factor, the conventional RL-based method without forgetting factor shows slow adaptation to changes in parameters. For this case, our RL-method is able to achieve 25% improvement in the convergence constant over conventional RL.

In Fig. 11, the impact of forgetting factor on the convergence of the RL framework is demonstrated. This figure shows the RL-based reward estimation error for the cooperative agent under two different forgetting factor values. As the forgetting factor increases from 0.01 to 0.1, the convergence speed of the RL framework has been improved. Hence, the forgetting factor controls the rate of adaptiveness to new conditions. However, a tradeoff exists between the rate of convergence and the accuracy of the solution. As can be seen, higher forgetting factors also lead to higher variances in the estimation error signal.

VI. Conclusion

Smart distribution systems with networked MGs in a cooperative setting can facilitate reliable power delivery to customers in future rural power grids. However, cooperatives can have incomplete knowledge of MG members' operational parameters due to data privacy and ownership concerns, which is an obstacle in the way of optimal decision making. Motivated by the shortcomings of model-based multiple MG power management in distribution systems with limited observability, this paper presents an adaptive RL-based methodology for bi-level power management of cooperatives consisting of multiple networked MGs.

We have shown that: 1) using the proposed decision method, a cooperative agent is able to accurately track the behavior of multiple networked MGs under incomplete knowledge of operation variables behind the PCCs. This can be used to indirectly control the response of participants in a price-based environment. 2) The proposed RL-based method is able to generalize from its past experiences to estimate optimal solutions in new situations without re-training from random initial conditions (i.e., fast response under evolving system conditions). This immensely speeds up the power management computational process. 3) The framework is shown to be adaptive against the changes happening to unobserved parameters that are excluded from cooperative agent's state set. The learning model has been tested and verified using extensive numerical scenarios. To summarize, the proposed decision model shows better adaptability, solution quality, and computational time compared to conventional centralized optimization methods.

The current RL-based decision model is limited to the power management of a single cooperative service provider with multiple MGs. However, in more realistic cases, there could also be multiple cooperative service providers in an interconnected rural area, which implies that the impact of cooperative service providers on each other and on the wholesale price could not be ignored. Hence, an optimal coordination scheme needs to be designed to enable collaboration among multiple entities. In future work, we will extend the proposed RL method to address this challenge.

Appendix
MG Optimal Power Management Formulation

A moving look-ahead decision window $[t, t+T]$ is defined using the latest estimations of solar and load power at different instants, where n is the MG index ($n \in \{1, \ldots, N\}$), i and j define the bus numbers for each MG ($\forall i, j \in \Omega_I$), and k denotes the line index ($\forall k \in \Omega_K$). It has decision vector $\boldsymbol{x_p} = (P^{DG}_{i,t,n}, P^{PCC}_{t,n}, P^{Ch}_{i,t,n}, P^{Dis}_{i,t,n})^\top$ and $\boldsymbol{x_q} = (Q^{DG}_{i,t,n}, Q^{PCC}_{t,n}, Q^{PV}_{i,t,n}, Q^{ESS}_{i,t,n})^\top$.

$$\min_{\boldsymbol{x_p}, \boldsymbol{x_q}} \sum_{t}^{T+t} \Big(-\lambda^R_{t,n} P^{PCC}_{t,n} + \lambda^F_{i,t,n} F_{i,t,n}\Big) \tag{16}$$

$$s.t. \;\; F_{i,t,n} = a_f \Big(P^{DG}_{i,t,n}\Big)^2 + b_f P^{DG}_{i,t,n} + c_f \tag{17}$$

$$\Big|P^{PCC}_{t,n}\Big| \le P^{PCC,M}_{t,n} \tag{18}$$

$$\Big|Q^{PCC}_{t,n}\Big| \le Q^{PCC,M}_{t,n} \tag{19}$$

$$0 \le P^{DG}_{i,t,n} \le P^{DG,M}_{i,n} \tag{20}$$

$$0 \le Q^{DG}_{i,t,n} \le Q^{DG,M}_{i,n} \tag{21}$$

$$\Big|P^{DG}_{i,t,n} - P^{DG}_{i,t-1,n}\Big| \le P^{DG,R}_{i,n} \tag{22}$$

$$P^{ij}_{t,n} = V^i_{t,n}\Big(V^i_{t,n} G^{ij}_n - V^j_{t,n}\Big(G^{ij}_n cos\Big(\Delta\theta^{ij}_{t,n}\Big) + B^{ij}_n sin\Big(\Delta\theta^{ij}_{t,n}\Big)\Big)\Big) \tag{23}$$

$$Q^{ij}_{t,n} = -V^i_{t,n}\Big(V^i_{t,n} B^{ij}_n + V^j_{t,n}\Big(G^{ij}_n cos\Big(\Delta\theta^{ij}_{t,n}\Big) - B^{ij}_n sin\Big(\Delta\theta^{ij}_{t,n}\Big)\Big)\Big) \tag{24}$$

$$\Big(P^{ij}_{t,n}\Big)^2 + \Big(Q^{ij}_{t,n}\Big)^2 \le \Big(L^{ij,M}_{t,n}\Big)^2 \tag{25}$$

$$\sum_{ij\in k}^{K} P^{ij}_{t,n} = \sum_{ji\in k}^{K} P^{ji}_{t,n} - p_{i,t,n} \tag{26}$$

$$\sum_{i,j\in k}^{K} Q^{ij}_{t,n} = \sum_{j,i\in k}^{K} Q^{ji}_{t,n} - q_{i,t,n} \tag{27}$$

$$p_{i,t,n} = P^{D,e}_{i,t,n} - P^{DG}_{i,t,n} - P^{PV,e}_{i,t,n} + P^{Ch}_{i,t,n} - P^{Dis}_{i,t,n} \tag{28}$$

$$P^D_{i,t,n} = P^{D,e}_{i,t,n} - \varepsilon^D_{i,t,n} \tag{29}$$

$$P^{PV}_{i,t,n} = P^{PV,e}_{i,t,n} - \varepsilon^{PV}_{i,t,n} \tag{30}$$

$$q_{i,t,n} = Q^D_{i,t,n} - Q^{DG}_{i,t,n} - Q^{PV}_{i,t,n} + Q^{ESS}_{i,t,n} \tag{31}$$

$$V^{PCC}_{t,n} = V^{PCC,E}_{t,n} \tag{32}$$

$$V^m_{i,n} \le V_{i,t,n} \le V^M_{i,n} \tag{33}$$

$$\Big|Q^{PV}_{i,t,n}\Big| \le Q^{PV,M}_{i,n} \tag{34}$$

$$SOC_{i,t,n} = SOC_{i,t-1,n} + \Delta t\Big(P^{Ch}_{i,t,n}\eta_{Ch} - P^{Dis}_{i,t,n}/\eta_{Dis}\Big)/E^{Cap}_{i,n} \tag{35}$$

$$SOC^m_{i,n} \le SOC_{i,t,n} \le SOC^M_{i,n} \tag{36}$$

$$0 \le P^{Ch}_{i,t,n} \le u^{Ch}_{i,t,n} P^{Ch,M}_{i,n} \tag{37}$$

$$0 \le P^{Dis}_{i,t,n} \le u^{Dis}_{i,t,n} P^{Dis,M}_{i,n} \tag{38}$$

$$0 \le u^{Ch}_{i,t,n} + u^{Dis}_{i,t,n} \le 1 \tag{39}$$

$$u^{Ch}_{i,t,n}, u^{Dis}_{i,t,n} \in \{0, 1\} \tag{40}$$

The objective function (16) minimizes each MG's total cost of operation, which is composed of two terms: the negative of revenue from power transfer with the cooperative agent and the cost of running local DGs. Here, $\lambda^F_{t,n}$ is the diesel generator fuel price in $\$/L$ adopted from [32]. The fuel consumption $F_{i,t,n}$ of diesel generator can be expressed as a quadratic polynomial function (17), with coefficients $a_f = 0.0001773\ L/kW^2$, $b_f = 0.1709\ L/kW$, and $c_f = 14.67L$ adopted from [33]. Constraints (18)-(19) describe the power exchange limit between the MG and the upstream distribution grid with the maximum active/reactive power exchange limits, $P^{PCC,M}_{t,n}, Q^{PCC,M}_{t,n}$. Constraints (20)-(21) ensure that the DG active/reactive power outputs, $P^{DG}_{i,t,n}/Q^{DG}_{i,t,n}$, are within the DG power capacity $P^{DG,M}_{i,n}, Q^{DG,M}_{i,n}$, and (22) enforces the maximum DG ramp limit, $P^{DG,R}_{i,n}$. Internal AC power flow model of the MG is considered here with the network topology constraints, with (23) and (24) determining the active and reactive power flows of each branch, where G^{ij} and B^{ij} are the corresponding real and imaginary parts of the bus admittance matrix, and $V^i_{t,n}$ and $\Delta\theta^{ij}_{t,n}$ are the nodal voltage magnitude and phase angle difference, respectively. Constraint (25) denotes the power flow limits for each branch. Equations (26)-(31) are the nodal active/reactive power balances at MG buses.

The difference between the predicted and actual PV/load values are modeled using Gaussian error variables as shown in equations (29) and (30), where $P_{i,t,n}^{D,e}$ denotes the estimated active load, and $P_{i,t,n}^{PV,e}$ is the estimated active power output of PV. Also, $\varepsilon_{i,t,n}^{D}, \varepsilon_{i,t,n}^{PV} \sim N(0,\sigma)$ denote the Gaussian estimation errors for active load and PV power, respectively. Constraint (32) sets the voltage at the PCC of the MG according to the estimated input voltage, $V_{t,n}^{PCC,E}$. Constraint (33) sets the limits for nodal bus voltage amplitude, $[V_{i,n}^{m}, V_{i,n}^{M}]$. PV reactive power output, $Q_{i,t,n}^{PV}$, is constrained by its maximum limit $Q_{i,n}^{PV,M}$ in (34). Operational ESS constraints are described by (35)-(40). Adopted from [34], constraint (35) determines the state of charge (SOC) of ESSs, $SOC_{i,t,n}$. The SOC and charging/discharging power of ESS, $P_{i,t,n}^{Ch}$, $P_{i,t,n}^{Dis}$, are constrained in (36)-(40). Here, $[SOC_{i,n}^{m}, SOC_{i,n}^{M}]$, $P_{i,n}^{Ch,M}$ and $P_{i,n}^{Dis,M}$ define the permissible range of SOC, and maximum charging and discharging power, with $u_{i,t,n}^{Ch}$ and $u_{i,t,n}^{Dis}$ denoting the charge/discharge binary indicator variables, and η_{Ch}/η_{Dis} representing the charging/discharging efficiency. $E_{i,n}^{Cap}$ denotes the maximum capacity of ESSs.

REFERENCES

[1] S. A. Arefifar, Y. A.-R. I. Mohamed, and T. H. M. EL-Fouly, "Optimum microgrid design for enhancing reliability and supply-security," *IEEE Trans. Smart Grid*, vol. 4, no. 3, pp. 1567–1575, Sep. 2013.

[2] K. Ubilla *et al.*, "Smart microgrids as a solution for rural electrification: Ensuring long-term sustainability through cadastre and business models," *IEEE Trans. Sustain. Energy*, vol. 5, no. 4, pp. 1310–1318, Oct. 2014.

[3] P. Chen, "A roadmap to the new rural electric cooperative business model," M.S. thesis, Nicholas School Environ., Duke Univ., Durham, NC, USA, 2017.

AQ3

[4] *Understanding Value is Critical to Microgrid Projects*. [Online]. Available: https://www.ncelectriccooperatives.com/who-we-are/spotlight/understanding-value-is-critical-to-microgrid-projects/

[5] R. Lasseter *et al.*, *Integration of Distributed Energy Resources. The CERTS Microgrid Concept*. Berkeley, CA, USA: Lawrence Berkeley Nat. Lab., 2002.

[6] M. Manbachi and M. Ordonez, "AMI-based energy management of islanded AC/DC microgrids utilizing energy conversion and optimization," *IEEE Trans. Smart Grid*, to be published.

[7] M. Nemati, M. Braun, and S. Tenbohlen, "Optimization of unit commitment and economic dispatch in microgrids based on genetic algorithm and mixed integer linear programming," *Appl. Energy*, vol. 210, pp. 944–963, Jan. 2018.

[8] N. Nikmehr and S. N. Ravadanegh, "Optimal power dispatch of multi-microgrids at future smart distribution grids," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1949–1957, Jul. 2015.

[9] S. A. Arefifar, M. Ordonez, and Y. A. I. Mohamed, "Energy management in multi-microgrid systems—Development and assessment," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 910–952, Mar. 2017.

[10] A. Ouammi, H. Dagdougui, and R. Sacile, "Optimal control of power flows and energy local storages in a network of microgrids modeled as a system of systems," *IEEE Trans. Control Syst. Tech.*, vol. 62, no. 4, pp. 2551–2559, Apr. 2015.

[11] Y. Zhang, L. Xie, and Q. Ding, "Interactive control of coupled microgrids for guaranteed system-wide small signal stability," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 1088–1096, Mar. 2016.

[12] Z. Wang, B. Chen, J. Wang, and J. Kim, "Decentralized energy management system for networked microgrids in grid-connected and islanded modes," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 1097–1105, Mar. 2016.

[13] Z. Wang, B. Chen, J. Wang, M. Begovic, and C. Chen, "Coordinated energy management of networked microgrids in distribution systems," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 45–53, Jan. 2015.

[14] P. Kou, D. Liang, and L. Gao, "Distributed EMPC of multiple microgrids for coordinated stochastic energy management," *Appl. Energy*, vol. 185, pp. 939–952, Jan. 2017.

[15] D. Gregoratti and J. Matamoros, "Distributed energy trading: The multiple-microgrid case," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2551–2559, Apr. 2015.

[16] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 2192–2203, Jun. 2018.

[17] E. Mocanu *et al.*, "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.

[18] *Guidelines for Smart Grid Cybersecurity Volume 2: Privacy and the Smart Grid*. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2014/NIST.IR.7628r1.pdf

[19] G. Mokryani, "Active distribution networks planning with integration of demand response," *Solar Energy*, vol. 122, pp. 1362–1370, Dec. 2015.

[20] M. Wahbah, T. H. M. El-Fouly, B. Zahawi, and S. Feng, "Hybrid beta-KDE model for solar irradiance probability density estimation," *IEEE Trans. Sustain. Energy*, to be published.

[21] H. Wu, M. Shahidehpour, Z. Li, and W. Tian, "Chance-constrained day-ahead scheduling in stochastic power system operation," *IEEE Trans. Power Syst.*, vol. 29, no. 4, pp. 1583–1591, Jul. 2014.

[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. London, U.K.: MIT Press, 2017.

[23] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 14–25, Feb. 2011.

[24] L. Busoniu, R. Babuska, B. D. Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, FL, USA: CRC Press, 2017.

[25] W. D. Smart and L. P. Kaelbling, "Practical reinforcement learning in continuous spaces," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2000, pp. 903–910.

[26] I. Houtzager, J. W. van Wingerden, and M. Verhaegen, "Recursive predictor-based subspace identification with application to the real-time closed-loop tracking of flutter," *IEEE Trans. Control Syst. Technol.*, vol. 20, no. 4, pp. 934–949, Jul. 2012.

[27] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Trans. Power Del.*, vol. 4, no. 2, pp. 1401–1407, Apr. 1989.

[28] R. S. Rao, K. Ravindra, K. Satish, and S. V. L. Narasimham, "Power loss minimization in distribution system using network reconfiguration in the presence of distributed generation," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 317–325, Feb. 2013.

[29] W. H. Kersting, "Radial distribution test feeder," in *Proc. IEEE Power Eng. Soc. Winter Meeting*, vol. 2, 2001, pp. 908–912.

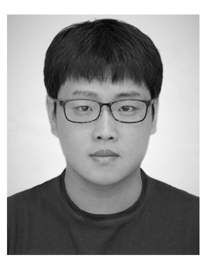
[30] *Wholesale Electricity and Natural Gas Market Data*. [Online]. Available: https://www.eia.gov/electricity/wholesale/#history

[31] B. Anthony, D. Kendrick, and A. Meeraus, "GAMS, a user's guide," *ACM SIGNUM Newslett.*, vol. 23, nos. 3–4, pp. 10–11, 1988.

[32] *Gasoline and Diesel Fuel Update*. [Online]. Available: https://www.eia.gov/petroleum/gasdiesel/

[33] S. A. Pourmousavi, M. H. Nehrir, and R. K. Sharma, "Multi-timescale power management for islanded microgrids including storage and demand response," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1185–1195, May 2015.

[34] Y. Xuan, J. Zhang, and Q. Zhang, "Combinational linear programming approach for daily optimal operation of customers with DG/ESS under TOU pricing for customer EMS software applications," in *Proc. IEEE Innov. Smart Grid Technol. Asia (ISGT Asia)*, 2018, pp. 300–305.



Qianzhi Zhang (S'17) received the B.S. degree in electrical and computer engineering from the Shandong University of Technology in 2012, and the M.S. degree in electrical and computer engineering from Arizona State University in 2015. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. From 2015 to 2016, he was a Research Engineer with Huadian Electric Power Research Institute, Hangzhou, China. His research interests include power distribution systems, microgrids, applications of distributed optimization, and machine learning in power systems.

Kaveh Dehghanpour (S'14–M'17) received the B.Sc. and M.S. degrees in electrical and computer engineering from the University of Tehran in 2011 and 2013, respectively, and the Ph.D. degree in electrical engineering from Montana State University in 2017. He is currently a Post-Doctoral Research Associate with Iowa State University. His research interest includes application of machine learning and data-driven techniques in power system monitoring and control.

Zhaoyu Wang (S'13–M'15) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiao Tong University in 2009 and 2012, respectively, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology in 2012 and 2015, respectively. He is the Harpole-Pentair Assistant Professor with Iowa State University. He was a Research Aid with Argonne National Laboratory in 2013 and an Electrical Engineer Intern with Corning Inc., in 2014. His research interests include power distribution systems, microgrids, renewable integration, power system resilience, and data-driven system modeling. He is the Principal Investigator for a multitude of projects focused on these topics and funded by the National Science Foundation, the Department of Energy, National Laboratories, PSERC, and Iowa Energy Center. He is the Secretary of the IEEE Power and Energy Society Award Subcommittee. He is an Editor of the IEEE TRANSACTIONS ON POWER SYSTEMS, the IEEE TRANSACTIONS ON SMART GRID, and IEEE PES Letters, and an Associate Editor of *IET Smart Grid*.

Qiuhua Huang (S'14–M'16) received the B.E. and M.Sc. degrees from the South China University of Technology in 2012 and 2009, respectively, and the Ph.D. degree in electrical engineering from Arizona State University in 2016. He is currently a Senior Power System Research Engineer with the Electricity Infrastructure Group, Pacific Northwest National Laboratory, USA. He has been a Principal Investigator, a Project Manager, and task lead in several DOE funded projects. His research interests include power system modeling, simulation and control, and application of advanced computing, and machine learning technologies in power systems. He was a recipient of the IEEE PES Prize Paper Award in 2019 and the R&D 100 Award in 2018. He is an Associate Editor of IEEE ACCESS and the *CSEE Journal of Power and Energy Systems*.