# Model-Free Real-Time Autonomous Energy Management for a Residential Multi-Carrier Energy System: A Deep Reinforcement Learning Approach

**Yujian Ye**[1,2*], **Dawei Qiu**[1], **Jonathan Ward**[2] and **Marcin Abram**[2]

[1]Department of Electrical and Electronic Engineering, Imperial College London, London, U.K.
[2]Fetch.ai, Cambridge, U.K.
{yujian.ye11, d.qiu15}@imperial.ac.uk, {yujian.ye, jonathan.ward, marcin.abram}@fetch.ai

## Abstract

The problem of real-time autonomous energy management is an application area that is receiving unprecedented attention from consumers, governments, academia and industry. This paper showcases the first application of deep reinforcement learning (DRL) to real-time autonomous energy management for a multi-carrier energy system. The proposed approach is tailored to align with the nature of the energy management problem by posing it in multi-dimensional continuous state and action spaces, in order to coordinate power flows between different energy devices, and to adequately capture the synergistic effect of couplings between different energy carriers. This fundamental contribution is a significant step forward from earlier approaches that only sought to control the power output of a single device and neglected the demand-supply coupling of different energy carriers. Case studies on a real-world scenario demonstrate that the proposed method significantly outperforms existing DRL methods as well as model-based control approaches in achieving the lowest energy cost and yielding a representation of energy management policies that adapt to system uncertainties.

## 1 Introduction

### 1.1 Background and Motivation

Energy systems worldwide are facing major flexibility challenges driven by the limited system ability to provide secure and economical supply-demand balance in face of the large-scale integration of renewable energy sources and their inherent variability. In this respect, it is increasingly recognized that integrated optimization and management of *multi-carrier energy systems* (MCES) represents a key opportunity to provide the required flexibility to support cost-effective evolution to smart low carbon future. This is enabled by their significant yet untapped potential to shift supply and demand across energy carriers by exploiting different forms of energy storage, compared to traditional energy systems whose carriers are operated independently [Mancarella, 2014].

At the residential level, with the increasing prevalence of distributed energy resources and advanced metering and communication infrastructure, end-users rely on *autonomous energy management systems* (AEMS) to actively control their generation, conversion, consumption and storage of energy in real-time, in order to reduce their demand placed onto the grid and their costs. Therein, substantial socio-environmental benefits emerge as a result of the reduced carbon emission.

An integrated solar photovoltaic (PV) and electricity energy storage (EES) system serves to compensate the mismatch between the time-varying and uncertain PV generation and electricity demand (ED) as well as making some revenue by injecting surplus PV to the electrical grid (EG). Meanwhile, the increasing adoption of electric vehicles (EV) by end-users is observed, driven by their ability to store electrical energy in their batteries, enabling the temporal decoupling between the absorption of energy from EG and its actual consumption for travelling purposes. Furthermore, coupling the operation of an electric heat pump (EHP), a gas boiler (GB) and a thermal energy storage (TES) introduces opportunities for an energy-shifting arbitrage between electricity and gas to supply electricity and heat to the end-users, which is particularly useful in face of intermittent RES. Furthermore, TES enables redistribution of heat demand (HD) and production across time which contributes to greater energy-shifting flexibility. In view of such complex interactions and inter-dependencies between different energy carriers as well as the multi-source system uncertainties, developing an effective AEMS plays an crucial role in uncovering the flexibility potential and harvesting the real-world benefits of the MCES.

### 1.2 Related Work

Energy management of the MCES has been traditionally addressed with *model-based control* approaches. A centralized operation cost minimization problem is solved to determine the optimal schedule of various kinds of controllable energy sources, loads and storage devices [Bozchalui *et al.*, 2012; Rastegar and Fotuhi-Firuzabad, 2015; Moghaddam *et al.*, 2016; Basit *et al.*, 2017]. This optimization problem generally requires full knowledge of the operational model and parameters of the MCES, and the energy usage schedules are determined using forecasted parameters, such as the energy demand, price patterns, EV user's commuting behavior and weather-dependent PV production. However, such de-

*Contact Author

terministic optimization models fail to account for the inherent uncertainties associated with these parameters. To this end, stochastic programming approaches are employed to schedule different energy devices [Pazouki *et al.*, 2014; Vahid-Pakdel *et al.*, 2017; Sedighizadeh *et al.*, 2018]. A scenario-based method is used to model the uncertainties. However, it poses significant challenges to identify appropriate probability distributions and construct a representative set of scenarios for all the involved uncertain parameters since they are influenced by various exogenous factors which are independent to the energy management strategy. Furthermore, the scale of the optimization problem increases drastically with the number of scenarios, resulting in significant computational burden [Conejo *et al.*, 2010]. An alternative approach consists in applying distributed control techniques. The alternating direction method of multipliers algorithm is adopted for energy usage scheduling of smart buildings and energy internet [Zhang *et al.*, 2017]. A consensus-based algorithm is employed in [Li *et al.*, 2016] for energy management of a combined heat and power system. However, the convergence to the optimal solution for both algorithms is only guaranteed assuming convex operating characteristics of the system and a very large number of iterations is generally necessary before a good measure of convergence can be obtained.

In contrast, reinforcement learning (RL) is a *model-free control approach* that consists of an agent (AEMS) gradually learning the optimal control policy by utilizing experiences acquired from its repeated interactions with the environment (MCES). RL also makes no assumption regarding the convexity of the operating characteristics of the MCES. In other words, the AEMS does not require prior information on the model dynamics of the MCES and considers the latter as a black box. In the big data era, RL can utilize the increasing volume of data collected from smart meters and perform successive interpretation of data to learn optimal management strategies and thereby coping with the uncertainties that are encapsulated in the data. The trained model can be deployed to deliver real-time control on timescales of milliseconds. Furthermore, RL enables a representation of the control actions to be constructed that generalizes to previously unseen situations. In this context, previous works have employed *Q-learning* (QL) for optimal demand response [Liang *et al.*, 2013; Wen *et al.*, 2015] and control of an integrated PV and EES system [Berlink *et al.*, 2015; Kim and Lim, 2018]. However, QL suffers severely from the *curse of dimensionality*. The discretization of both state and action spaces may distort the feedback that the agent receives regarding the influence of its actions on the environment and adversely affect the feasible action space, resulting in sub-optimal policies. This challenge is aggravated in the setting of the examined problem, since both state of the environment (e.g. PV output, the energy content of the EES) and agent's actions (e.g. charging / discharging schedule of EES) are not only continuous but also multi-dimensional.

In view of these limitations, more recently, the *deep Q network* (DQN) method is applied to determine home energy management decisions considering a heating, ventilation and air conditioning system [Wei *et al.*, 2017], shiftable loads [Mocanu *et al.*, 2019], EV [Wu *et al.*, 2018; Wan *et al.*, 2019] and an integrated renewable energy and EES system [Chen and Su, 2018]. DQN employs a deep neural network (DNN) to approximate the Q-value function. Despite the generalization capability of DQN to multi-dimensional continuous state space, its performs sub-optimally in problems with continuous action spaces, because the employed DNN is trained to produce discrete Q-value estimates rather than continuous actions [Lillicrap and *et al.*, 2016]. For instance, only seven discrete power levels are assumed for the EV charging in [Wan *et al.*, 2019] and the action of the EES is limited to three options: fully charging, fully discharging or idle in [Chen and Su, 2018]. This substantially limits the flexibility potential of energy storage devices and thus hinders the effectiveness of DQN in addressing the examined problem, since the energy management decisions are multi-dimensional and continuous. To this end, the *deep policy gradient* (DPG) method is adopted in [Mocanu *et al.*, 2019; Ye *et al.*, 2019]. DPG also employs a DNN, but instead of estimating the Q-value function, it directly estimates the probability of taking an action at a specific state. However, the underlying energy management actions considered in [Mocanu *et al.*, 2019] are limited to discrete on / off status of different loads while their actual energy consumption schedules are determined by solving a cost minimization problem, which implies that the employed management strategy is not model-free and discrete in nature. Furthermore, DPG generally suffers from high variance in its gradient estimates which results in slow convergence [Silver *et al.*, 2014].

### 1.3 Contributions

This paper attempts to fill the knowledge gap and address the fundamental limitations of previous approaches through the following novel contributions:

- A novel autonomous energy management strategy is proposed for a residential MCES by applying the prioritized deep deterministic policy gradient (PDDPG) method. This approach is model-free and requires neither the knowledge of the system modeling of the MCES nor any forecasted exogenous information, as opposed to traditional model-based approaches. To the best of the authors' knowledge, and according to the up-to-date literature review conducted in [Mason and Grijalva, 2019], this is the first time that this approach has been used to address the optimal energy management problem of an MCES.

- In contrast with earlier works which all target a single and isolated energy carrier and only sought to control the power output of a single energy device, the energy management of a MCES constitutes a more challenging task as it necessitates monitoring and managing the activity of each device in the MCES so as to adequately capture the synergistic effects of the couplings between different energy carriers.

- Furthermore, as opposed to previous works which largely simplify the energy management problem employing discrete control RL method, the proposed approach conforms to the nature of the energy management problem by setting it up in multi-dimensional continuous state and action spaces and properly accounting for the effect of non-convex operating characteristics of the MCES.

- The value of the proposed real-time energy manage-

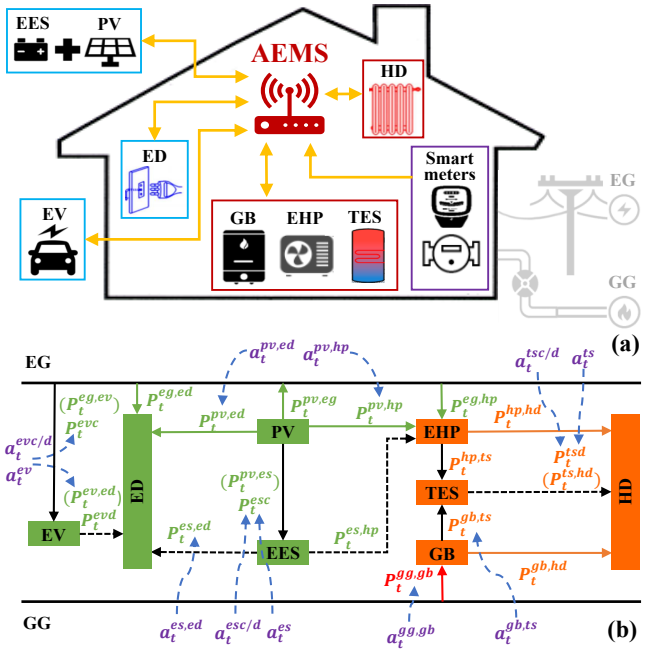


Figure 1: (a) Schematic representation of the MCES and (b) connections and power input / output of energy devices. The black solid and dashed arrows highlight mutually exclusive power flows.

ment strategy is validated through numerous numerical experiments in a real-world scenario accounting for uncertainties stemmed from the supply / demand sides of the MCES.

## 2 Energy Management as an MDP

A finite *Markov decision process* (MDP) with discrete time step is applied to formulate the energy management problem. The time interval between two adjacent steps is one hour. The AEMS constitutes the agent, while the residential MCES represents the environment, featuring an integrated electricity, heat and gas systems, as depicted in Figure 1 (a). The electricity system is composed of the EG, a PV, an EES and an EV. The heat system consists of an EHP, a GB and a TES. The MCES is also connected to the gas grid (GG) for the operation of GB. The objective of AEMS is to optimally manage the energy usage schedules of all devices in each carrier of the MCES so as to minimize the total cost for the end-user.

**State.** The state $s_t$ at time step $t$ is defined as a 12-dimensional vector $s_t = [E_t, t, \overline{P}_t^{ev}, E_t^{tr}, P_t^d, \lambda_t, P_t^{pv}]$, where $E_t = [E_t^{es}, E_t^{ev}, E_t^{ts}]$ represent the energy contents of the EES, EV and TES, these are endogenous state features which are affected by agent's actions; $t$ denotes the step identifier; $\overline{P}_t^{ev}$ indicates the maximum charging / discharging limit[1] of EV; $E_t^{tr}$ represents the electrical energy requirements[2] of EV for travelling purposes; $P_t^d = [P_t^{ed}, P_t^{hd}]$ represent ED and HD; $\lambda_t = [\lambda_t^{e-}, \lambda_t^{e+}, \lambda_t^g]$ represent electricity buy / sell prices and gas price; and $P_t^{pv}$ denotes the PV production. These are exogenous state features which are decoupled from agent's actions and are characterized by inherent variability and uncertainty. Note that, our approach takes into account the consumer preference by incorporating the EV user's commuting behavior as exogenous state features.

**Action.** Given the state $s_t$, the actions $a_t$ of the AEMS at time step $t$ is defined as a 11-dimensional vector $a_t = [a_t^{esc/d}, a_t^{evc/d}, a_t^{tsc/d}, a_t^{es}, a_t^{ev}, a_t^{ts}, a_t^{gg,gb}, a_t^{gb,ts}, a_t^{pv,hp}, a_t^{pv,ed}, a_t^{es,ed}]$. As illustrated in Figure 1 (b), the blue dashed arrows indicate the action that manages the respective power input / output for each device of the MCES.

Specifically, $a_t^{esc/d}$, $a_t^{evc/d}$ and $a_t^{tsc/d} \in \{0, 1\}$ represent the charging (1) / discharging (0) status of the EES, EV and TES, these actions represent the inherent non-convex operating characteristics of energy storage devices [Ye *et al.*, 2014], ensuring that they operate exclusively either in the charging or discharging mode; $a_t^{es}$, $a_t^{ev}$ and $a_t^{ts} \in [0, 1]$ represent the charging / discharging power of EES, EV and TES as a percentage of their maximum limits $\overline{P}^{es}$, $\overline{P}_t^{ev}$ and $\overline{P}^{ts}$; $a_t^{gg,gb} \in [0, 1]$ represents the gas input of GB as a percentage of its maximum limit $\overline{P}^{gb}$; $a_t^{gb,ts} \in [0, 1]$ represents the charging power of TES from GB as a percentage of the total charging power of TES $P_t^{tsc}$; $a_t^{pv,hp}$ and $a_t^{pv,ed} \in [0, 1]$ represent the power flow from PV to EHP and to ED both as a percentage of $P_t^{pv}$; and $a_t^{es,ed} \in [0, 1]$ represent the discharging power of EES to supply ED as a percentage of its total discharging power $P_t^{esd}$.

**State transition.** The state transition from $s_t$ to $s_{t+1}$ is governed by a function: $s_{t+1} = F(s_t, a_t, \omega_t)$. The transition may be not only affected by the action $a_t$ but also influenced by the randomness $\omega_t$ existed in some state features. In the examined problem, the transitions for $\overline{P}_t^{ev}$, $E_t^{tr}$, $P_t^d$, $\lambda_t$ and $P_t^{pv}$ are subject to variability and uncertainties. Identifying probability distributions to accurately capture such randomness can be very challenging since they are affected by many factors, such as EV user's commuting behaviour, pricing process of the utility, and the weather conditions. To resolve this issue, a model-free approach is proposed to learn the transition for such features from real-world data-set using machine learning techniques. On the other hand, the transitions for $E_t^{es}$, $E_t^{ev}$ and $E_t^{ts}$ are directly affected by energy management actions. After executing actions $a_t$, the state transitions of these features as well as the resultant power flows indicated in the MCES (Figure 1 (b)) can be derived according to the operational constraints that characterize the MCES. For space limitation reasons, these derivations are not presented here.

**Reward.** The reward $r_t$ resultant from the energy management decisions $a_t$ is set to be equal to the negative operation cost $C_t$ of the MCES as given by:

$$C_t = \lambda_t^{e-} P_t^{eg,(ed+hp+ev)} + \lambda_t^g P_t^{gg,gb} - \lambda_t^{e+} P_t^{pv,eg} \quad (1)$$

where the three terms represent, respectively, the cost of purchasing electricity from EG, purchasing gas from GG, and the revenue from from selling excess PV production to EG.

[1] $\overline{P}_t^{ev}$ is equal to a fixed charging / discharging limit when EV is parked at home (assuming home-charging) and 0 otherwise.

[2] $E_t^{tr}$ is equal to a fixed energy consumption level (dependent to the distance travelled) when EV is traveling and 0 otherwise.

**Performance and value functions.** At each time step, the agent employs a policy $\pi$ to interact with the MDP and emit a trajectory of states, actions and rewards: $s_1, a_1, r_1, s_2, a_2, r_2, ...$ over $\mathcal{S} \times \mathcal{A} \times \mathbb{R}$. The return $R_t = \sum_{l=t}^{T} \gamma^{(l-t)} r_l$ is the discounted reward where $\gamma \in [0, 1]$ is the discount factor. The agents' goal through RL is to construct a policy that maximizes the cumulative discounted reward, denoted by the performance function $J(\pi) = \mathbb{E}[R_1|\pi] = \mathbb{E}_{s\sim\rho^\pi, a\sim\pi}[r]$, where $\rho^\pi$ denotes the discounted state distribution. The Q-value function $Q_\pi(s, a) = \mathbb{E}[R_1|s_1 = s, a_1 = a; \pi]$ forms an estimation of the discounted reward.

## 3 Proposed Energy Management Strategy

As discussed in Section 1.2, despite the wide popularity of adopting QL and DQN for energy management problems in the most recent smart grid literature, these approaches suffer from the curse of dimensionality to some extent, since they necessitate discretization of the state and / or the action spaces. However, the discretization of the state space may distort the feedback that the agent receives regarding the impact of its actions on the environment while the discretization of the action space may adversely affect the feasible action space, resulting in sub-optimal policies. Furthermore, although DPG can be used for continuous control, the gradient estimator of DPG generally suffers from low sampling efficiency and high variance in its gradient estimates which results in slow convergence [Konda and Tsitsiklis, 2000].

Aiming at addressing these limitations, the PDDPG method [Ye *et al.*, 2020] features an *actor-critic* architecture and employs two DNNs for different purposes [Lillicrap and *et al.*, 2016]. The critic network $Q_\theta$ takes as input a state $s_t$ and action $a_t$ and outputs a scalar estimate of the Q-value function $Q_\theta(s_t, a_t)$. The actor network $\mu_\phi$ takes as input a state $s_t$ and implements the policy improvement task which updates the policy with respect to the estimated Q-value function and outputs a continuous action $\mu_\phi(s_t)$. Learning the Q-value function in addition to the policy serves to significantly reduce the variance in the gradient estimates compared to DPG and consequently promises better convergence properties.

Concerning policy improvement, the common approach adopted in QL and DQN is a greedy maximization of the Q-value function. However, greedy policy improvement tends to be intractable in multi-dimensional continuous action spaces as it necessitates maximizing the Q-value function globally at every time step. Instead, the proposed method employs the actor network $\mu$ to generate an action $\mu_\phi(s_{t+1})$ for the next state. The critic network then implements the policy evaluation task, appraising the policy by producing an estimate of the Q-value function with TD learning. Rather than globally maximizing $Q_\theta(s_t, a_t)$, the critic calculates gradients $\nabla_a Q_\theta(s_t, a_t)$ which indicate directions of change of action resulting in higher estimated Q-values. These gradients are computed via back-propagation through the critic, which is more computational efficient than solving an optimization problem in continuous action space.

Analogous to DQN, PDDPG incorporates target network and the experience replay [Mnih and *et al.*, 2015] as mechanisms to stabilize the training process. In the former, an online and a target network are used to separate the Q-value update and the target Q-value evaluation. The weights of the target networks are updated by having them slowly tracking the online networks to constrain the target values to change slowly so as to improve the stability of the learning process. In the latter, the sequentially generated training experiences are stored in a replay buffer and sampled to train the DNNs, diminishing the temporal correlations existed in the replayed experiences. To further enhance the sampling efficiency, the *prioritized experience replay* [Schaul *et al.*, 2016] method is employed which prioritizes learning from experiences corresponding to higher absolute TD error which promises both improved policy and faster learning speed. When prioritized sampling a minibatch of $N$ transitions $\{(s_n, a_n, r_n, s_{n+1})\}_{n=1}^{N}$, the actor is updated by applying the *policy gradient theorem* [Silver *et al.*, 2014]:

$$\nabla_\phi J(\mu_\phi) = N^{-1} \sum_n \nabla_a Q_\theta(s_n, \mu_\phi(s_n)) \nabla_\phi \mu_\phi(s_n) \quad (2)$$

Then, by defining the absolute TD error $|\delta_n|$, the critic is updated by minimizing the loss function $\mathcal{L}_\theta$,

$$|\delta_n| = |r_n + \gamma Q_{\theta'}(s_{n+1}, \mu'_\phi(s_{n+1})) - Q_\theta(s_n, a_n)| \quad (3)$$

$$\mathcal{L}_\theta = N^{-1} \sum_n W_n \delta_n^2 \quad (4)$$

where $W_n$ represents the weighting factors related to the probability of sampling experience $n$ based on $|\delta_n|$.

## 4 Case Studies

### 4.1 Experiment Setup and Implementation

A real-world scenario developed by the UK government [Pudjianto *et al.*, 2013] is used to train and evaluate the proposed energy management strategy. The data is provided over a yearly horizon and hourly resolution. The electricity buy price follows the time-of-use tariff structure provided in [Hydro, 2019], partitioned into summer and winter periods (Figure 2). The sell price is set as the UK feed-in tariff [Ofgem, 2019]. The gas price is provided by a major UK gas supplier [E.ON, 2019]. The average driving patterns of the EV is taken from [UK Department for Transport, 2018]. Based on which, the EV is assumed to make two journeys per day, each is defined by a start time, end time and electrical energy requirement. The EV is assumed connected to EG during the period between the end of their second and the start of their first journey. The technical parameters of EES, TES, EHP and GB are taken from [Pudjianto *et al.*, 2013].

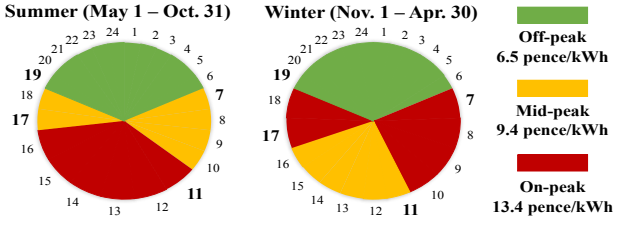


Figure 2: Time-of-use tariff structure.

The one-year data-set is split into training, validation and test sets. In each month, the first 20 days are used for training, and the remaining days are selected for performance evaluation. The values of most of the hyperparameters associated with the design and training of the neural networks and the design of the PER have been gathered from [Lillicrap and *et al.*, 2016] and [Schaul *et al.*, 2016], respectively. We found that the discount rate and the mini-batch size have the greatest impact on the algorithm performance. To tune these hyperparameters, we split the training data and used the first 15 days for training and the last 5 days for validation.

To demonstrate the value of PDDPG, we compare it with DQN and DPG. Their implementations are briefly discussed as follows. In order to apply DQN, we discretize all the action dimensions, apart from the first three, in three integer values representing 0%, 50%, and 100% energy usage. It should be noted that although the action space can be discretized with a higher granularity, it leads to an exponential growth of the number of actions, and the resultant DQN is impractical to train. In order to model continuous control, we represent the probability distribution of agent's action with a normal distribution, and predict the mean and variance of it with a DNN.

## 4.2 Performance Evaluation

In this section, the proposed PDDPG approach is evaluated and compared with several benchmark solutions, including DQN, DPG, as well as three model-based approaches. The first one employs a "theoretical" optimal controller which minimizes the daily energy cost, assuming full knowledge of the model and parameters of the MCES and perfect forecast of the uncertain parameters. This controller formalizes the problem as a mixed-integer linear program (MILP), the optimal solution of which can be regarded as a lower bound on the cost, indicating how far from the optimum the model-free DRL controllers are. The second one is an hourly ("myopic") MILP which solves an hourly energy cost minimization problem that neglects the time-coupling operating characteristics of all energy storage devices. The third one resorts to model predictive control (MPC) [Kou *et al.*, 2015]. At each time step, the MPC computes the control action by solving a cost minimization over a rolling horizon of 8 hours, and only the first element of the obtained control sequence is implemented. An LSTM network is used to forecast the exogenous system parameters. We train each DRL method for $2 \times 10^4$ episodes for 10 different random seeds. Each episode represents a random day from the training set and consists of 24 time steps. We assess the quality of the energy management strategy every 200 episodes during training by evaluating it on the test set. Figure 3 illustrates the average daily cost over the 125 test days for the examined three DRL methods with 10 seeds. The mean and the standard deviation of the average daily cost over the 10 seeds are illustrated through the solid lines and the shaded areas, respectively, in Figure 3. The cumulative daily energy costs of the test 125 days under PDDPG and all examined baseline methods are presented in Figure 4.

As illustrated in Figure 3, among the model-free DRL approaches, PDDPG improves its policy in a stable fashion and eventually only PDDPG manages to converge to a near-optimum solution with a decreasing standard deviation.

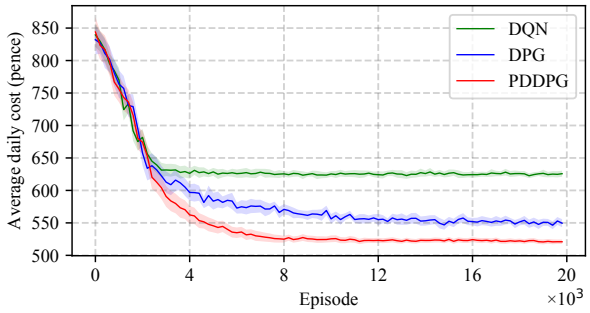


Figure 3: Average daily cost evaluated over the test set for the examined DRL methods with 10 different random seeds.

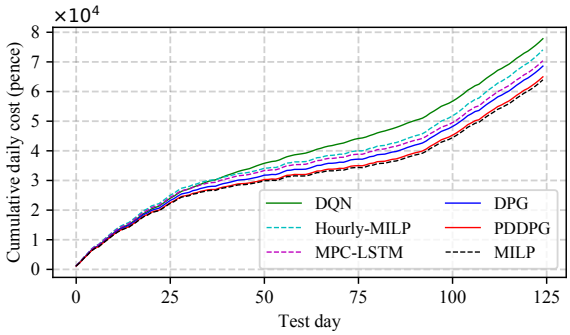


Figure 4: Cumulative daily costs of PDDPG and all the baseline methods over the 125 test days.

PDDPG significantly outperforms benchmark DRL methods, achieving the lowest average daily cost 520.95 pence and the smallest standard deviation 3.48 pence at convergence. In relative terms, PDDPG achieves 16.75% / 5.20% lower average daily cost and 25.32% / 44.10% lower standard deviation over DQN / DPG, respectively. PDDPG and DPG both outperform DQN towards the objective of cost saving. This is attributed to their ability to model multi-dimensional continuous action space, in contrast to the naïve discretization approach employed in DQN. DPG and PDDPG allow the AEMS to represent more accurate information from the entire action space and thus discover more cost-effective management strategy by exploiting it. Furthermore, PDDPG exhibits advantageous convergence properties compared to DPG in terms of the obtained average daily cost and learning stability. Its superior performance could be explained by i) PDDPG implements a critic, which estimates the Q-value function and appraises each action that the agent takes. In contrast, DPG lacks the policy evaluation step, resulting in high variance in its gradient estimates and ii) PDDPG incorporates the PER mechanism which more frequently replays experiences corresponding to higher TD-error, improving the learning performance.

As depicted in Figure 4, the costs obtained by the four benchmark approaches, DPG (blue solid line), MPC-LSTM (magenta dashed line), Hourly-MILP (cyan dashed line), and DQN (green solid line) are 7.24%, 10.08%, 15.83%, and 21.71% higher than the theoretical optimum, respectively. Among the model-based baselines, neglecting the time-coupling operating characteristics of EES, TES and EV, hourly MILP corresponds to the highest average daily cost.

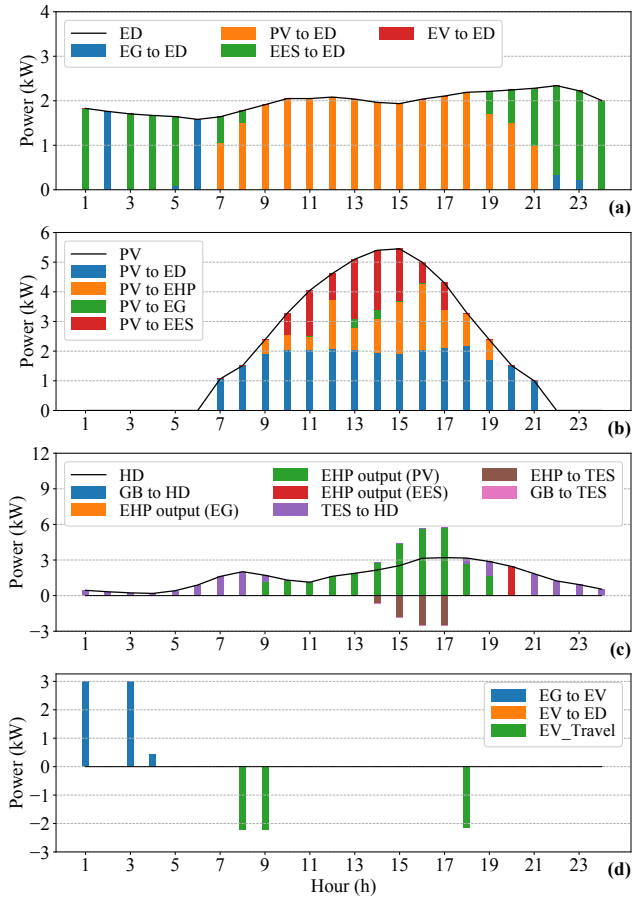


Figure 5: (a) balancing of ED, (b) usage of PV generation, (c) balancing of HD and (d) EV charging/discharging pattern for the examined summer day under PDDPG.

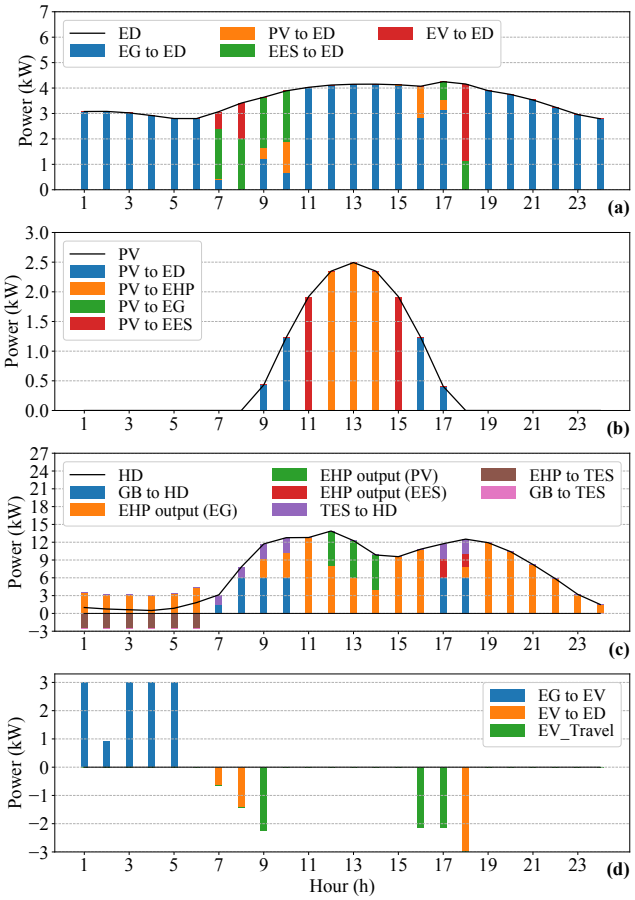


Figure 6: (a) balancing of ED, (b) usage of PV generation, (c) balancing of HD and d) EV charging/discharging pattern for the examined winter day under PDDPG.

The cost under MPC-LSTM is second highest as a result of accumulated forecasting errors of the exogenous parameters. In comparison, PDDPG (red solid line) achieves an average cost that is only 1.84% higher than the theoretical optimum (black dashed line), outperforming significantly both model-free DRL and model-based baselines.

To further investigate the performance of PDDPG in coping with the MCES uncertainties, we deploy the trained model and analyse the obtained energy usage schedules of the MCES for a typical summer and winter day selected in the 125 test days, as shown in Figures 5 and 6, respectively. The summer day (Figure 5) is characterized by abundant PV generation and small HD. PDDPG learns to supply the majority of ED from PV and EES discharge; as such, the end-user only imports energy from EG during a few off-peak hours, leading to a significant cost saving. The majority of the peak PV generation is stored in EES (which discharges at night to supply ED when PV production is low or none) instead of selling to EG because electricity buy price at night is still higher than the sell price. EV is charged at lowest-priced early morning hours to fulfill its energy requirement to travel. Furthermore, PDDPG learns to exploit the flexibility offered by multiple supply sources and storage devices so as to supply HD at zero energy cost. Specifically, EHP is controlled to take advantage of the abundant PV generation and the EES discharge to supply HD; additionally, the operation of TES is coupled with EHP which enables a redistribution of HD across time, diminishing the need of EHP to source any electricity from EG.

The winter day (Figure 6), on the other hand, is characterized by scarce PV generation and large HD. As a result, the overall energy import from EG and GG are both increased. Nevertheless, PDDPG can adaptively adjust to the winter condition towards the cost-saving objective by making very efficient use of the limited PV generation during the mid-day periods to supply ED and HD (through EHP and TES), and also charge a significant amount of PV to EES, which later discharges at peak ED and HD periods where the electricity buy price is also at its highest. Additionally, apart from fulfilling EV's traveling energy requirement, more energy is charged to the EV battery during off-peak periods and discharged to supply ED before the first and after the second EV journey when the buy price adopts its mid-peak value and the PV generation is absent, significantly shaving the ED peak. Furthermore, during the periods where the electricity buy price is relatively low, and driven by the considerably higher conversion efficiency of EHP (compared to GB), PDDPG learns to source the supply of HD from EHP. During the periods where electricity buy price is at its peak, it makes

efficient use of i) GB to convert gas to heat and ii) TES to charge from EHP during the off-peak periods and discharge to supply HD at these periods.

### 4.3 Value of Continuous Energy Management

This section analyzes in more depth the physical significance of PDDPG to model continuous actions in the context of the examined problem by comparing to the DQN method. Figure 7 illustrates the energy scheduling decisions of the MCES obtained using DQN for the examined summer day.

Due to the discretization of actions $a_t^{es}$ and $a_t^{pv,hp}$, the end-user imports more energy from EG, leading to a higher cost of 238.86 pence (PDDPG: 26.00 pence) for supplying ED. DQN also makes inefficient use of the bountiful PV production compared to PDDPG with only a small amount of PV supplied to EHP and charged to EES, but a significant amount sold to EG. Consequently, to meet the majority of HD, EHP is required to purchase a significant amount of electricity from EG, leading to a sharp cost increase of 80.50 pence (PDDGP: 0) for supplying HD. Furthermore, TES can only charge / discharge at peak HD periods resultant from the discretisation of action $a_t^{ts}$, limiting its flexibility in redistributing HD in time. Lastly, due to the discretisation of action $a_t^{ev}$, EV needs to charge for 3 hours at the maximum limit (3kW) to fulfill its daily traveling energy requirement (6.6kWh), as indicated in Figure 5(d), contributing to a higher cost of 58.50 pence (PDDGP: 41.75 pence) for EV charging.

Overall, although the MCES obtains a higher revenue from selling PV surplus to EG, the daily energy cost of DQN (239.96 pence) is approximately 3.68 times as high as the one of PDDPG (65.22 pence). Furthermore, the end-user's total energy demand placed on the grids of DQN (42.19kWh) is around 4.05 times as high as the one of PDDPG (10.42kWh). It can be concluded that learning in discrete action space may result in sub-optimal management strategies where the non-convex operating characteristics of different storage devices and complex interactions between different energy carriers must be taken into account. On the other hand, PDDPG preserves all the relevant information concerning the entire action space, and thus is capable of learning more cost-effective management strategies.

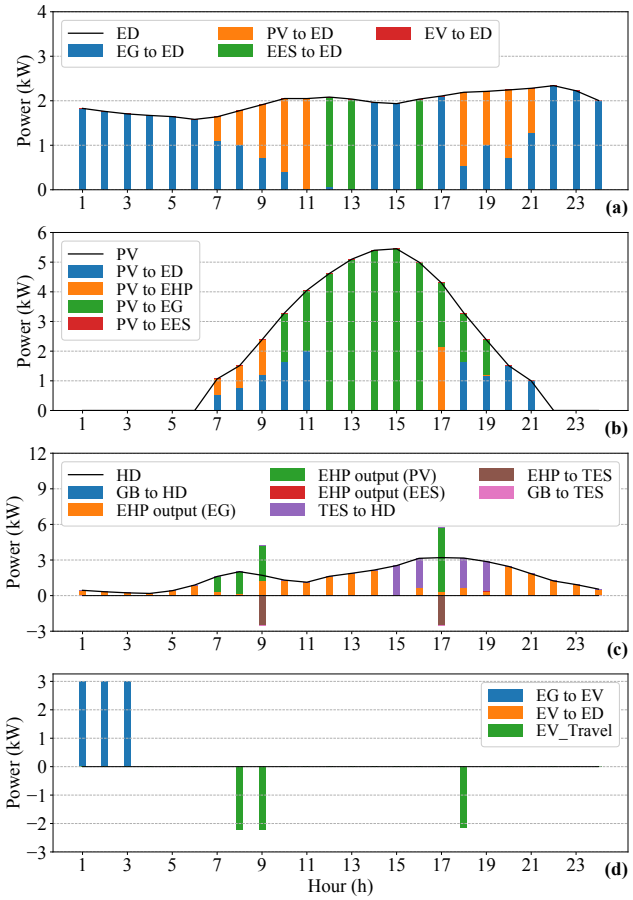


Figure 7: (a) balancing of ED, (b) usage of PV generation, (c) balancing of HD and (d) EV charging/discharging pattern for the examined summer day under DQN.

## 5 Conclusions

This paper showcases the first application of PDDPG to real-time autonomous energy management for a MCES. The problem is formalised as a MDP which takes into account the uncertainties stemming from both the supply and demand sides of the MCES. The proposed approach is model-free and data-driven, which does not rely on knowledge of the system modeling of the MCES and any forecasted exogenous information. The proposed approach respects the nature of the examined problem by posing it multi-dimensional continuous state and action spaces, enabling the AEMS to receive accurate feedback regarding the impact of its energy management strategy on the status of the MCES, and devise more cost-effective strategies by exploiting the entire action domain, also accounting for the effect of non-convex operating characteristics of the MCES.

Case studies on a real-world scenario have provided numerous new and valuable insights around the value of the proposed energy management strategy. Quantitative results have demonstrated that PDDPG manages to converge to a near-optimum solution and achieves a significantly higher cost savings for the MCES compared with existing model-free DRL as well as model-based baselines. Furthermore, PDDPG enables representation of real-time energy management policies to be constructed that can adaptively cope with system uncertainties towards the cost-saving objective and generalize well to previously unseen situations.

## References

[Basit *et al.*, 2017] Abdul Basit, Guftaar A. S. Sidhu, Anzar Mahmood, and Feifei Gao. Efficient and Autonomous Energy Management Techniques for the Future Smart Homes. *IEEE Trans. on Smart Grid*, 8(2):917–926, 2017.

[Berlink *et al.*, 2015] Heider Berlink, Nelson Kagan, and Anna H. R. Costa. Intelligent Decision-Making for Smart Home Energy Management. *J. Intell. Robot. Syst.*, 80(1):331–354, 2015.

[Bozchalui *et al.*, 2012] Mohammad C. Bozchalui, Syed A. Hashmi, Hussin Hassen, Claudio A. Canizares, and Kankar Bhattacharya. Optimal Operation of Residential Energy Hubs

in Smart Grids. *IEEE Trans. on Smart Grid*, 3(4):1755–1766, 2012.

[Chen and Su, 2018] Tao Chen and Wencong Su. Local Energy Trading Behavior Modeling with Deep Reinforcement Learning. *IEEE Access*, 6:62806–62814, 2018.

[Conejo *et al.*, 2010] Antonio J Conejo, Miguel Carrión, Juan M Morales, et al. *Decision Making Under Uncertainty in Electricity Markets*, volume 1. Springer, 2010.

[E.ON, 2019] E.ON. Fix Online v27 Gas Tariff, 2019.

[Hydro, 2019] London Hydro. Smart Meters and Time-of-Use Rates, 2019.

[Kim and Lim, 2018] Sunyong Kim and Hyuk Lim. Reinforcement Learning Based Energy Management Algorithm for Smart Energy Buildings. *Energies*, 11(8):1–19, 2018.

[Konda and Tsitsiklis, 2000] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Proc. NIPS*, pages 1008–1014, CO, USA, 2000.

[Kou *et al.*, 2015] Peng Kou, Deliang Liang, Lin Gao, and Feng Gao. Stochastic coordination of plug-in electric vehicles and wind turbines in microgrid: A model predictive control approach. *IEEE Trans. on Smart Grid*, 7(3):1537–1551, 2015.

[Li *et al.*, 2016] Yu-Shuai Li, Hua-Guang Zhang, Bo-Nan Huang, and Fei Teng. Distributed Optimal Economic Dispatch Based on Multi-Agent System Framework in Combined Heat and Power Systems. *Appl. Sci.*, 6(10):308, 2016.

[Liang *et al.*, 2013] Yong Liang, Long He, Xinyu Cao, and Zuo-Jun Shen. Stochastic Control for Smart Grid Users with Flexible Demand. *IEEE Trans. on Smart Grid*, 4(4):2296–2308, 2013.

[Lillicrap and *et al.*, 2016] Timothy P. Lillicrap and *et al.* Continuous Control with Deep Reinforcement Learning. In *Proc. ICLR*, pages 1–14, San Juan, US, 2016.

[Mancarella, 2014] Pierluigi Mancarella. MES (Multi-Energy Systems): An Overview of Concepts and Evaluation Models. *Energy*, 65:1–17, 2014.

[Mason and Grijalva, 2019] Karl Mason and Santiago Grijalva. A Review of Reinforcement Learning for Autonomous Building Energy Management. *Comput. Electr. Eng.*, 78:300–312, 2019.

[Mnih and *et al.*, 2015] Volodymyr Mnih and *et al.* Human-Level Control Through Deep Reinforcement Learning. *Nature*, 518(7540):529–533, 2015.

[Mocanu *et al.*, 2019] Elena Mocanu, Decebal C. Mocanu, Phuong H. Nguyen, Antonio Liotta, Michael E. Webber, Madeleine Gibescu, and Johannes G. Slootweg. On-Line Building Energy Optimization Using Deep Reinforcement Learning. *IEEE Trans. on Smart Grid*, 10(4):3698–3708, 2019.

[Moghaddam *et al.*, 2016] Iman G. Moghaddam, Mohsen Saniei, and Elaheh Mashhour. A Comprehensive Model for Self-Scheduling an Energy Hub to Supply Cooling, Heating and Electrical Demands of a Building. *Energy*, 94:157–170, 2016.

[Ofgem, 2019] Ofgem. Feed-In Tariff (FIT) Rates, 2019.

[Pazouki *et al.*, 2014] Samaneh Pazouki, Mahmoud-Reza Haghifam, and Albert Moser. Uncertainty Modeling in Optimal Operation of Energy Hub in Presence of Wind, Storage and Demand Response. *Int. J. Elec. Power.*, 61:335–345, 2014.

[Pudjianto *et al.*, 2013] Danny Pudjianto, Marko Aunedi, Predrag Djapic, and Goran Strbac. Whole-Systems Assessment of the Value of Energy Storage in Low-Carbon Electricity Systems. *IEEE Trans. on Smart Grid*, 5(2):1098–1109, 2013.

[Rastegar and Fotuhi-Firuzabad, 2015] Mohammad Rastegar and Mahmud Fotuhi-Firuzabad. Load Management in a Residential Energy Hub with Renewable Distributed Energy Resources. *Energy and Buildings*, 107:234–242, 2015.

[Schaul *et al.*, 2016] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay. In *Pro. ICLR*, pages 1–21, San Juan, US, 2016.

[Sedighizadeh *et al.*, 2018] Mostafa Sedighizadeh, Masoud Esmaili, and Nahid Mohammadkhani. Stochastic Multi-Objective Energy Management in Residential Microgrids with Combined Cooling, Heating, and Power Units Considering Battery Energy Storage Systems and Plug-in Hybrid Electric Vehicles. *J. Clean. Prod.*, 195:301–317, 2018.

[Silver *et al.*, 2014] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic Policy Gradient Algorithms. In *Proc. ICML*, pages 1–9, Beijing, China, 2014.

[UK Department for Transport, 2018] UK Department for Transport. National Travel Survey, 2018.

[Vahid-Pakdel *et al.*, 2017] MJ Vahid-Pakdel, Sayyad Nojavan, B Mohammadi-Ivatloo, and Kazem Zare. Stochastic Optimization of Energy Hub Operation with Consideration of Thermal Energy Market and Demand Response. *Energy Convers. Manag.*, 145:117–128, 2017.

[Wan *et al.*, 2019] Zhiqiang Wan, Hepeng Li, Haibo He, and Danil Prokhorov. Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning. *IEEE Trans. on Smart Grid*, 10(5):5246–5257, 2019.

[Wei *et al.*, 2017] Tianshu Wei, Yanzhi Wang, and Qi Zhu. Deep Reinforcement Learning for Building HVAC Control. In *Proc. DAC*, pages 1–7, Austin, USA, 2017.

[Wen *et al.*, 2015] Zheng Wen, Daniel O'Neill, and Hamid Maei. Optimal Demand Response Using Device-Based Reinforcement Learning. *IEEE Trans. on Smart Grid*, 6(5):2312–2324, 2015.

[Wu *et al.*, 2018] Di Wu, Guillaume Rabusseau, Vincent Francois-lavet, Doina Precup, and Benoit Boulet. Optimizing Home Energy Management and Electric Vehicle Charging with Reinforcement Learning. In *Proc. ALA*, pages 1–8, Stockholm, Sweden, 2018.

[Ye *et al.*, 2014] Yujian Ye, Dimitrios Papadaskalopoulos, and Goran Strbac. Factoring flexible demand non-convexities in electricity markets. *IEEE Trans. on Power Syst.*, 30(4):2090–2099, 2014.

[Ye *et al.*, 2019] Yujian Ye, Dawei Qiu, Jing Li, and Goran Strbac. Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: A novel multi-agent deep reinforcement learning approach. *IEEE Access*, 7:130515–130529, 2019.

[Ye *et al.*, 2020] Yujian Ye, Dawei Qiu, Mingyang Sun, Dimitrios Papadaskalopoulos, and Goran Strbac. Deep reinforcement learning for strategic bidding in electricity markets. *IEEE Trans. on Smart Grid*, 11(2):1343–1355, 2020.

[Zhang *et al.*, 2017] Huaguang Zhang, Yushuai Li, David Wenzhong Gao, and Jianguo Zhou. Distributed optimal energy management for energy internet. *IEEE Trans. Ind. Informat.*, 13(6):3081–3097, 2017.