Does Explicit Prediction Matter in Energy Management Based on Deep Reinforcement Learning?

Zhaoming Qin, Huaying Zhang, Yuzhou Zhao, Hong Xie, and Junwei Cao, *Senior Member, IEEE*

***Abstract*—As a model-free optimization and decision-making method, deep reinforcement learning (DRL) has been widely applied to the filed of energy management in energy Internet. While, some DRL-based energy management schemes also incorporate the prediction module used by the traditional model-based methods, which seems to be unnecessary and even adverse. In this work, we present the standard DRL-based energy management scheme with and without prediction. Then, these two schemes are compared in the unified energy management framework. The simulation results demonstrate that the energy management scheme without prediction is superior over the scheme with prediction. This work intends to rectify the misuse of DRL methods in the field of energy management.**

***Index Terms*—Deep reinforcement learning, energy management, prediction, recurrent neural network.**

I. INTRODUCTION

As alternative to conventional fossil fuels, there have been large-scale integration of the renewable energy sources (RESs) including solar power and wind power into power system [1]. Although RESs have advantages including sustainable and environmental friendly, it is intractable to conduct energy management with the penetration of high-proportional RESs due to the uncertainty and stochasticity of renewable generation output [2]. Moreover, the challenges for energy management are further exacerbated by the varying power demands and fluctuating electricity prices [3], [4]. Therefore, it is of great importance to develop the advanced energy management scheme to accommodates various disturbances from RESs, power demands and electricity prices.

Tremendous research effort has been dedicated in developing the *model-based* energy management schemes [5], [6], [7], [8]. A typical model-based approach is model predictive control (MPC), in which control signals are decided by solving an optimization problem with a finite time horizon, following a receding horizon approach. The formulated optimization problem generally relies on the access to full knowledge of the system model and parameters. Put differently, the optimal energy management scheduling is estimated using forecasted exogenous parameters, including the power demands, electricity prices and weather-dependent PV production. As a result, the performance of the consequent energy management schemes is significantly dependent to the accuracy of the employed system model and the forecasting method. Therefore, massive advanced predictive models and approaches have been developed [9], [10]. A novel hybrid modeling method using both deep neural networks (DNNs) and stochastic differential equations is proposed to To obtain accurate power models of photovoltaic panels and loads in [9]. The long short-term memory recurrent neural network (RNN) is employed to address the short-term residential load forecasting issue in [10].

By contrast, the model-free energy management schemes do not require the explicit system model and the predictive exogenous parameters, regarded as a potential alternative to model-based schemes [11], [12]. For example, the model-free reinforcement learning (RL) can gradually learn the optimal or near-optimal strategies by utilizing experiences collected from massive interactions with the environment, without a priori knowledge of the environment. Moreover, with the booming development of deep learning (DL) technologies, the deep reinforcement learning (DRL) has attracted great attention [13]. The DRL can be viewed as the combination of DL and RL. The powerful representation capability of DNNs enables DRL to address the continuous and high-dimensional state spaces and action spaces [14]. An energy management algorithm based on deep deterministic policy gradient (DDPG) is proposed to minimize the energy cost of smart home in [15]. Authors in [16] develop an vectorized DRL algorithm based on advantage actor-critic (A2C) to reduce the operation cost and improve user experience without some users' private information.

Although the DRL-based methods do not rely on the predictive models and parameters, some works still integrate the forecasting methods into model-free DRL, such as [17], [18]. Authors in [17] use feedforward DNNs to predict the future electricity prices which are served as the part of observation in DRL. Similarly, authors in [18] establish a price forecasting model using multilayer perceptron (MLP) used for the decision-making of DDPG algorithm. Prediction is indeed a dimensionality reduction processing of the original information, in this sense, the information received by the agent of DRL is not complete. Consequently, the powerful feature extraction capability cannot fully utilized. Despite the essential role of prediction in model-based methods, adding

This work was supported by Science and Technology Project of China Southern Power Grid (090000KK52190169/SZKJXM2019669).

Z. Qin is with the Department of Automation, Tsinghua University, Beijing, P. R. China.

H. Zhang, Y. Zhao and H. Xie are with New Smart City High-quality Power Supply Joint Laboratory of China Southern Power Grid (Shenzhen Power Supply CO., LTD).

J. Cao is with Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, P. R. China.

Corresponding author: Junwei Cao, email: jcao@tsinghua.edu.cn.

prediction to model-free methods may undermine the control effect.

In this work, we investigate the performance comparison of energy management schemes with and without explicit prediction. First, we formulate the general energy management problem of a microgrid as a Markov decision process (MDP). Second, we realize the energy management scheme with prediction, by training the forecasting models with SL and the policy with DRL respectively. Third, we implement the energy management scheme without prediction, by training the end-to-end policy network consisting of MLP and RNN. Finally, we conduct the simulation experiments to compare the effects of these two schemes. The main contributions of this paper can be summarized as follows.

- We investigate the effects of prediction in the DRL-based energy management scheme. To the best of our knowledge, this is the first paper to make a rigorous comparison between the DRL-based scheme with and without prediction.
- We establish the unified energy management framework under which the comparison between DRL-based scheme with and without prediction can be conducted fairly.
- Simulation results demonstrate that the DRL-based scheme without prediction outperforms over the scheme with prediction. Moreover, we intuitively explain how the prediction undermines the control effect of DRL.

II. Problem Formulation

A. System Decription

In this work, wo consider a general energy management problem of a microgrid. As shown in Fig. 1, the microgrid is comprised of RESs, non-adjustable loads, battery energy storage devices (BESs) and energy management system (EMS). The RESs could be solar panels and wind generators. The power demands of non-adjustable loads must be satisfied completely without delay. We suppose that the microgrid hourly operates in discrete time, i.e., $t \in \{0, 1, \ldots, T\}$ where T is the time horizon. Each time step begins at the beginning of the current hour and expires at the beginning of the next hour. For example, the period from 0:00 to 1:00 is time step 1, the period from 1:00 to 2:00 is time step 2, and so on. Moreover, the electricity price is announced hourly by the utility grid.

At the beginning of each hour, the EMS observes the renewable generation output and power demand during last hour, receives current state of charge (SOC) from BESs and hour-ahead electricity price from the utility grid. Then, the EMS determines the charging/discharging power of BESs. After the decision of EMS, if energy shortage occurs during this hour, the microgrid will purchase appropriate energy from the utility gird; while the excess energy will be abandoned.

B. Markov Decision Process Formulation

In this work, the energy management scheme is formulated as a MDP. A general MDP can be described as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R)$, where $\mathcal{S}, \mathcal{A}, \mathcal{P}, R$ are the state space, action space, transition dynamics and reward function. At each time step t, the agent observe a state $\mathbf{s}_t$ from the state space $\mathcal{S}$, and selects an action a_t from action space $\mathcal{A}$. After performing action a_t, state $\mathbf{s}_t$ transitions to state $\mathbf{s+1}_t$ with probability distribution $\mathcal{P}(\mathbf{s}_t, a_t)$. Additionally, the agent receives a scalar reward $r_t = R(\mathbf{s}_t, a_t)$. The goal of the MDP is to maximize the cumulative discount reward $R_0 = \sum_{t=0}^{T} \gamma^t r_t$. In the remainder of the section, the state, action, dynamics and reward will be specified.

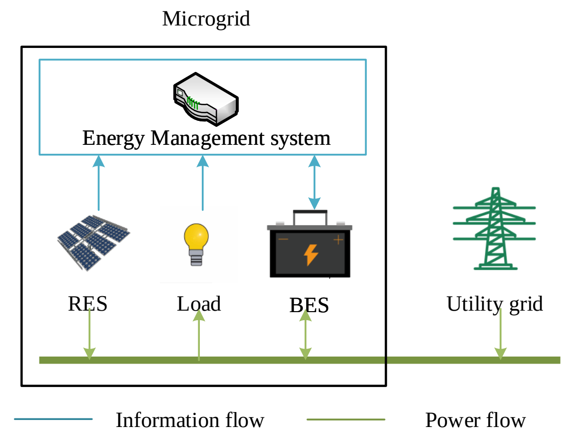


Fig. 1. Illustration of considered microgrid.

1) State: At each time step t, the state available to EMS includes the renewable generation output and power demand at last time step, the current SOC of BESs and the hour-ahead electricity price.

$$\boldsymbol{s}_t = [b_t, g_{t-1}, d_{t-1}, p_t], \tag{1}$$

where b_t, g_{t-1}, d_{t-1} and p_t are the SOC of BESs, renewable generation output, nonshiftable power demand and electricity price, respectively.

2) Action: The action a_t denotes the charging/discharging power at time step t, constrained by

$$-d_{max} \le a_t \le c_{max}, \tag{2}$$

where d_{max} and c_{max} are the maximum discharging power and charging power, respectively.

3) Dynamics: Normally, the dynamics of generation output, power demand and electricity price are difficult to describe precisely. In this work, the real historical data is directly utilized, including power data [19] and price data [20].

The dynamics of SOC is presented as follows

$$b_{t+1} = f(b_t, a_t), \tag{3}$$

where $f(\cdot)$ denote the transition function of SOC with respect to current SOC and charging/discharging power.

The power balance is guaranteed by purchasing energy from the utility grid.

$$e_t = \begin{cases} d_t - g_t + b_t, & \text{if } d_t - g_t + b_t > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where e_t is the power drawn from the utility grid at time step t.

4) Reward:

$$r_t = -p_t \cdot e_t - g\left(b_t, a_t\right), \tag{5}$$

where $g(\cdot)$ is the degradation cost function of BESs with respect to current SOC and charging/discharging power.

III. DRL-Based Energy Management Scheme

In this section, we present the DRL-based scheme with and without prediction. First, SL is applied to train a RNN to conduct the prediction. Then, the DRL-based scheme with the prediction module is realized. Finally, the end-to-end DRL-based scheme without prediction is proposed.

A. SL for Prediction

The target of SL is to learn a function f parameterized by ϕ such that $y = f(x; \varphi)$. Here, x and y denote the input and the label, respectively. Under our scenario, we intend to predict the future renewable generation output, power demand and hour-ahead price, which belongs to time series prediction problem. Considering the outperformance of RNNs with the processing of temporal sequence, in this work, gated recurrent units (GRUs), a gating mechanism in RNNs, are employed to represent the forecasting models.

The training process of SL is shown in Algorithm 1. The input x_t could be generation output g_t, power demand d_t and electricity price p_t. The GRUs are trained with back-propagation such that the mean square error (MSE) between the outputs of GRUs and the target values is minimized.

Algorithm 1 SL for k-step prediction

Initialize parameter φ_k
for epoch $= 1$ to N **do**
  for time step $t = 0$ to T **do**
    $\hat{x}_{t+k}, h_{t+1} = GRU(x_t, h_t; \varphi_k)$
    $d\varphi_k \leftarrow d\varphi_k + \nabla_{\varphi_k}\left(\hat{x}_{t+k} - x_{t+k}\right)^2$
  Perform update of φ_k using $d\varphi_k$

B. DRL-Based Scheme with Prediction

Given the trained forecasting models, in this subsection, the DRL-based scheme with prediction is presented.

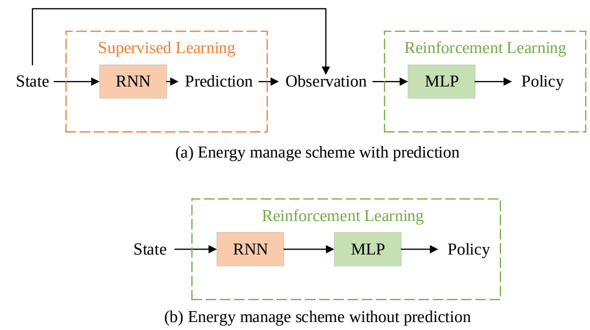
First, the observation of DRL is the concatenation of state and prediction as follows,

$$\boldsymbol{o}_t = \boldsymbol{s}_t \cup \left[\hat{g}_t, \hat{d}_t, \hat{p}_{t+1}, \ldots, \hat{g}_{t+k}, \hat{d}_{t+k}, \hat{p}_{t+k}\right], \tag{6}$$

where the expanded part is predicted by the forecasting models.

Then, the policy of DRL is trained by PPO algorithm [21] which maintains the actor network $\pi\left(a_t|\mathbf{o}_t;\theta\right)$ parameterized by θ and the critic network $V(\mathbf{o}_t;\phi)$ parameterized by ϕ. The parameters of actor are updated by minimizing following loss function,

$$\mathcal{L}_a(\theta) = \mathbb{E}_t\left[\min\left(w_t\hat{A}_t, \text{clip}\left(w_t, 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right], \tag{7}$$



(a) Energy manage scheme with prediction
(b) Energy manage scheme without prediction

Fig. 2. Overview of energy management schemes.

where $\hat{A}_t$ denotes the advantage calculated as

$$\hat{A}_t = \sum_{t'=t}^{T}(\gamma\lambda)^{t'-t}\left(-V(\mathbf{o}_{t'};\phi) + r_t + \gamma V\left(\mathbf{o}_{t'+1};\phi\right)\right), \tag{8}$$

and w_t is the probability ratio defined as

$$w_t = \frac{\pi(a_t|\mathbf{o}_t;\theta)}{\pi(a_t|\mathbf{o}_t;\theta_{\text{old}})}. \tag{9}$$

Accordingly, the parameter of critic is updated by minimizing

$$\mathcal{L}_c(\phi) = \mathbb{E}_t\left[\left(V\left(\mathbf{o}_t;\phi\right) - \hat{R}_t\right)^2\right]. \tag{10}$$

The detail of the training is presented in Algorithm 2.

Algorithm 2 PPO for scheme with prediction

Initialize parameter θ and ϕ for actor and critic
Load parameter φ for GRUs
for episode $= 0$ to N **do**
  $h_0 \leftarrow 0$
  for $t = 0$ to T **do**
    Perform prediction $x_t, h_{t+1} = GRU(\mathbf{s}_t, h_t; \varphi)$
    $\mathbf{o}_t \leftarrow [\mathbf{s}_t, x_t]$
    $\mathcal{P} \leftarrow \pi(\cdot|\mathbf{o}_t;\theta)$, $v_t = V(\mathbf{o}_t;\phi)$
    Sample action a_t according to distribution $\mathcal{P}$
    Execute action a_t and observe $\mathbf{s}_{t+1}$
    Compute the probability $p_t^{old} \leftarrow \mathcal{P}(a_t)$
  $\hat{A}_T \leftarrow 0, v_T \leftarrow 0$
  for $t = T-1$ to 0 **do**
    $\hat{R}_t \leftarrow \gamma\lambda\hat{A}_{t+1} + r_t + \gamma v_{t+1}$
    $\hat{A}_t \leftarrow \hat{R}_t - v_t$
  for $k = 1$ to K **do**
    $\mathcal{L}_a \leftarrow 0, \mathcal{L}_c \leftarrow 0$
    for $t = 0$ to $T-1$ **do**
      $w_t \leftarrow \pi(a_t|\mathbf{o}_t;\theta)/p_t^{old}$
      $\mathcal{L}_a += \min\left(w_t\hat{A}_t, \text{clip}\left(w_t, 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)$
      $\mathcal{L}_c += \left(V(\mathbf{o}_t;\phi) - \hat{R}_t\right)^2$
    Update θ and ϕ with gradient $\nabla_\theta\mathcal{L}_a$ and $\nabla_\phi\mathcal{L}_c$

C. DRL-Based Scheme without Prediction

As shown in Fig. 2, under the scheme with prediction, the RNN and MLP are trained with SL and RL, respectively. While the scheme without prediction performs end-to-end training. Put differently, the networks of actor and critic are comprised of RNN and MLP, rather than only MLP. In this sense, during the training process, the RNN could automatically learn appropriate parameters such that the most important information at previous time steps could be captured for decision.

The detail of DRL-based scheme without prediction is express in Algorithm 3. The state s_t is directly served as input to generate the policy and estimated value function. Simultaneously, the hidden states for actor and critic are generated for the next calculation.

Algorithm 3 PPO for scheme without prediction

Initialize parameter θ and ϕ for actor and critic
for episode $= 0$ to N **do**
  $h_0^\pi \leftarrow 0, h_0^V \leftarrow 0$
  for $t = 0$ to T **do**
    $\mathcal{P}, h_{t+1}^\pi = \pi(\mathbf{s}_t, h_t^\pi; \theta)$
    Sample action a_t according to distribution $\mathcal{P}$
    $p_t^{old} \leftarrow \mathcal{P}(a_t)$
    $v_t, h_{t+1}^V = V(\mathbf{s}_t, h_t^V; \phi)$
    Execute action a_t and observe $\mathbf{s}_{t+1}$
  $\hat{A}_T \leftarrow 0, v_T \leftarrow 0$
  for $t = T-1$ to 0 **do**
    $\hat{R}_t \leftarrow \gamma\lambda\hat{A}_{t+1} + r_t + \gamma v_{t+1}$
    $\hat{A}_t \leftarrow \hat{R}_t - v_t$
  for $k = 1$ to K **do**
    $\mathcal{L}_a \leftarrow 0, \mathcal{L}_c \leftarrow 0, h_0^\pi \leftarrow 0, h_0^V \leftarrow 0$
    for $t = 0$ to $T-1$ **do**
      $\mathcal{P}, h_{t+1}^\pi = \pi(\mathbf{s}_t, h_t^\pi; \theta)$
      $V_t, h_{t+1}^V = V(\mathbf{s}_t, h_t^V; \phi)$
      $w_t \leftarrow \mathcal{P}(a_t)/p_t^{old}$
      $\mathcal{L}_a += \min\left(w_t\hat{A}_t, \text{clip}\left(w_t, 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)$
      $\mathcal{L}_c += (V_t - \hat{R}_t)^2$
    Update θ and ϕ with gradient $\nabla_\theta\mathcal{L}_a$ and $\nabla_\phi\mathcal{L}_c$

IV. Performance Evaluation

In this section, the performances of DRL-based energy management scheme with and without prediction are compared. First, the simulation environment settings and algorithmic implementation are provided. Then, the performances of prediction used for energy management scheme are evaluated. Finally, the simulation results of comparisons between two schemes and corresponding explanation are given.

A. Environment Setup

We consider the energy management problem during one day, such that the time horizon T is 24. The transition functions and cost function are specified as follows [16].

$$f\left(b,a\right) = \begin{cases} b + \frac{\eta_c}{C}a, & \text{if } a \geq 0, \\ b + \frac{1}{C\eta_d}a, & \text{otherwise,} \end{cases} \tag{11}$$

$$g\left(b,a\right) = \begin{cases} \lambda_1|a|, & \text{if } b < 0.5, \\ \lambda_2|a|, & \text{otherwise,} \end{cases} \tag{12}$$

where C, η_c and η_d denotes the capacity, charging and discharging efficiency coefficients of BESs, λ_1 and λ_2 are the maximum and minimum degradation cost per kWh, corresponding to low SOC and high SOC, respectively. The parameter settings are provided in Table I.

TABLE I
Environment and Algorithmic Parameter Settings

Parameter	Value	Parameter	Value	Parameter	Value
d_{max}	400 kW	c_{max}	400 kW	C	2000 kWh
η_d	0.95	η_c	0.95	λ_1	0.013
λ_2	0.005	ϵ	0.2	K	3

We employ 10 parallel threads to interact with the environment. We use the historical data from 2015-01-05 to 2018-12-17 for the SL training, while the data from 2018-12-18 to 2020-03-23 is used for the test of prediction performance. For the training of DRL, the discount factor γ is set to be 0.95, The learning rate of actor and critic is set to be 3×10^{-4} and 1×10^{-3}, respectively. Other important algorithmic parameters are shown in Table I.

B. Performance of Prediction

During training process of SL, The MSE losses of the prediction for renewable generation output, power demand and electricity price are shown in Fig. 3, Fig. 4 and Fig. 5, respectively. It can be observed from these three figures that the MSE losses rapidly decrease and eventually converge, which demonstrates the stable training of RNNs.

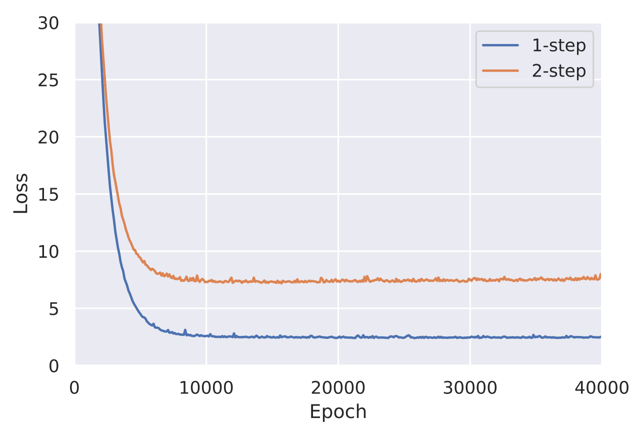


Fig. 3. Loss curve of prediction for renewable generation output.

The prediction effects during two days are shown in Fig. 6, Fig. 7 and Fig. 8. It can be observed from these three figures that the 1-step prediction is more accurate than 2-step prediction, which is also revealed by the loss curves during training.

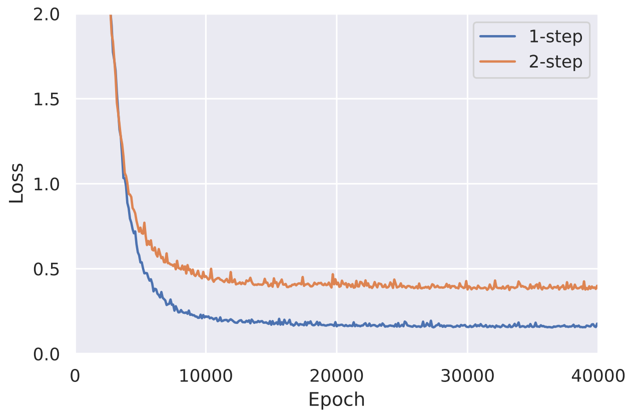


Fig. 4. Loss curve of prediction for power demand.

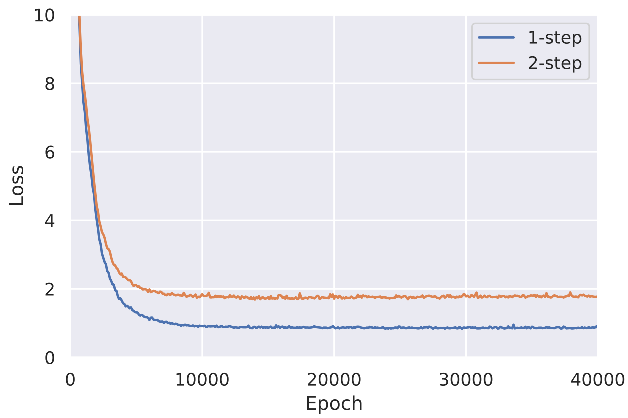


Fig. 5. Loss curve of prediction for electricity price.

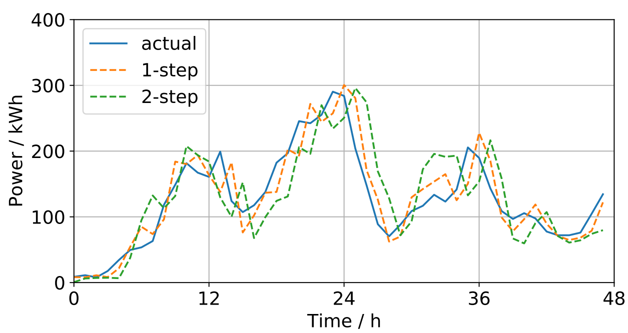


Fig. 6. Prediction for renewable generation output.

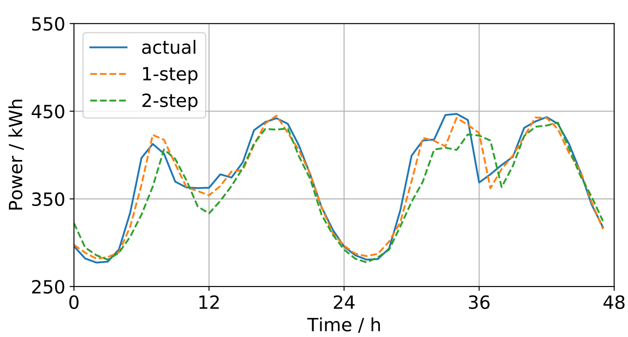


Fig. 7. Prediction for power demand.

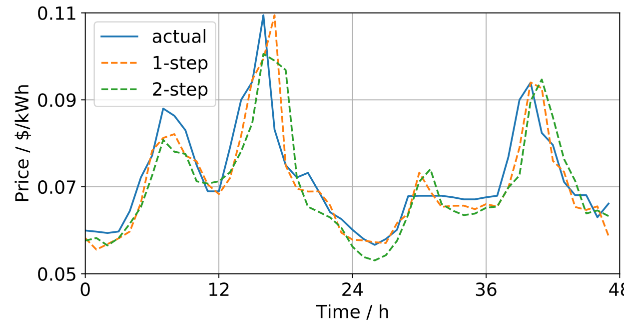


Fig. 8. Prediction for electricity price.

We adopt two metrics to evaluate the performances of prediction: mean absolute percentage error (MAPE) and root-mean-square error (RMSE). The evaluation results are shown in Table II.

TABLE II
TEST PERFORMANCE FOR PREDICTION

	1-step		2-step	
Metrics	MAPE	RMSE	MAPE	RMSE
Renewable generation output	17.7%	31.0	31.0%	47.0
Power demand	3.0%	13.3	5.7%	22.8
Electricity Price	8.2%	0.0046	11.4%	0.0058

C. Performance of Energy Management Schemes

The evaluate the performances of these two schemes during the training process of PPO, we depict the mean episode reward ($R = \sum_{t=0}^{T} r_t$) in Fig. 9. We can see that the energy management scheme without prediction has higher episode reward than scheme with prediction.

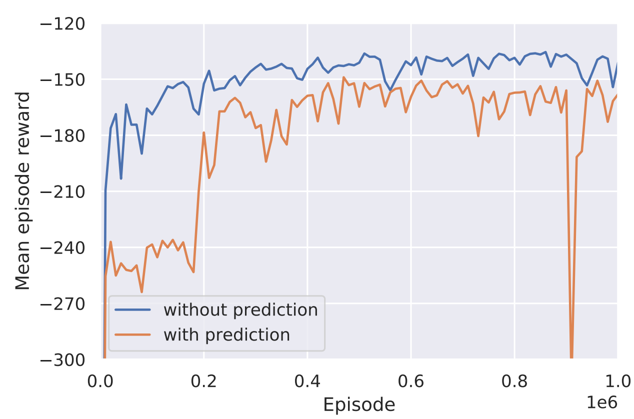


Fig. 9. Training curve.

Under the energy management scheme without prediction, the curves of charging/discharging power of BESs with electricity price are shown in Fig. 10. One can observe that the scheme without explicit prediction learns to charge when the electricity price is low and to discharge when the price is on-peak. These charging/discharging patterns demonstrate the DRL-based energy management scheme can accommodates the varying electricity price without explicit prediction.

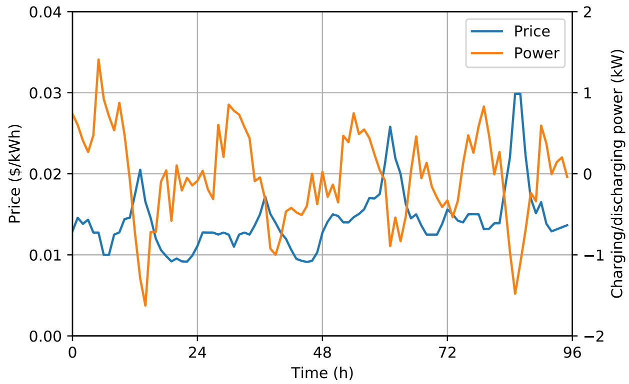


Fig. 10. Charging/discharging power of the scheme without explicit prediction over 4 consecutive days.

V. Conclusion

In this paper, we investigate whether the prediction matters in DRL-based energy management scheme. We present the standard energy management scheme with and without explicit prediction. The former is implemented with both SL and DRL, while the latter is directly implemented with end-to-end DRL. The simulation results demonstrate that end-to-end DRL enables the EMS to learn better control policies without explicit prediction sessions. This work can clarify the misuse and misunderstanding for DRL methods in the field of energy management.

References

[1] K. Rahbar, C. C. Chai, and R. Zhang, "Energy cooperation optimization in microgrids with renewable energy integration," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1482–1493, 2018.

[2] E. Du, N. Zhang, B.-M. Hodge, Q. Wang, C. Kang, B. Kroposki, and Q. Xia, "The role of concentrating solar power toward high renewable energy penetrated power systems," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6630–6641, 2018.

[3] M. Tan, S. Yuan, S. Li, Y. Su, H. Li, and F. He, "Ultra-short-term industrial power demand forecasting using lstm based hybrid ensemble learning," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2937–2948, 2020.

[4] J. Yang, J. Zhao, F. Wen, and Z. Dong, "A model of customizing electricity retail prices based on load profile clustering analysis," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3374–3386, 2019.

[5] X. Huang, S. H. Hong, and Y. Li, "Hour-ahead price based energy management scheme for industrial facilities," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 2886–2898, 2017.

[6] A. Saad, T. Youssef, A. T. Elsayed, A. Amin, O. H. Abdalla, and O. Mohammed, "Data-centric hierarchical distributed model predictive control for smart grid energy management," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4086–4098, 2019.

[7] P. Kou, D. Liang, and L. Gao, "Distributed empc of multiple microgrids for coordinated stochastic energy management," *Applied energy*, vol. 185, pp. 939–952, 2017.

[8] E. R. Stephens, D. B. Smith, and A. Mahanti, "Game theoretic model predictive control for distributed energy demand-side management," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1394–1402, 2015.

[9] Y. Qin, H. Hua, and J. Cao, "Stochastic optimal control scheme for battery lifetime extension in islanded microgrid via a novel modeling approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4467–4475, 2019.

[10] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[11] Y. Du and F. Li, "Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1066–1076, 2020.

[12] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6629–6639, 2019.

[13] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[14] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, and X. Guan, "A review of deep reinforcement learning for smart building energy management," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[15] L. Yu, W. Xie, D. Xie, Y. Zou, D. Zhang, Z. Sun, L. Zhang, Y. Zhang, and T. Jiang, "Deep reinforcement learning for smart home energy management," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2751–2762, 2020.

[16] Z. Qin, D. Liu, H. Hua, and J. Cao, "Privacy preserving load control of residential microgrid via deep reinforcement learning," *IEEE Transactions on Smart Grid*, pp. 1–1, 2021.

[17] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3201–3211, 2020.

[18] B. Liu, M. Akcakaya, and T. E. Mcdermott, "Automated control of transactive hvacs in energy distribution systems," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2462–2471, 2021.

[19] Pecan Street Database. [Online]. Available: http://www.pecanstreet.org/.

[20] ENGIE Resources. [Online]. Available: http://www..engieresources.com/.

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