ИНФОРМАТИКА

UDC 519.217
MSC 90C40

Microgrid control for renewable energy sources based on deep reinforcement learning and numerical optimization approaches*

A. Yu. Zhadan, H. Wu, P. S. Kudin, Y. Zhang, O. L. Petrosian

St. Petersburg State University,
7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Zhadan A. Yu., Wu H., Kudin P. S., Zhang Y., Petrosian O. L. Microgrid control for renewable energy sources based on deep reinforcement learning and numerical optimization approaches. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2023, vol. 19, iss. 3, pp. 391–402. https://doi.org/10.21638/11701/spbu10.2023.307

Optimal scheduling of battery energy storage system plays crucial part in distributed energy system. As a data driven method, deep reinforcement learning does not require system knowledge of dynamic system, present optimal solution for nonlinear optimization problem. In this research, financial cost of energy consumption reduced by scheduling battery energy using deep reinforcement learning method (RL). Reinforcement learning can adapt to equipment parameter changes and noise in the data, while mixed-integer linear programming (MILP) requires high accuracy in forecasting power generation and demand, accurate equipment parameters to achieve good performance, and high computational cost for large-scale industrial applications. Based on this, it can be assumed that deep RL based solution is capable of outperform classic deterministic optimization model MILP. This study compares four state-of-the-art RL algorithms for the battery power plant control problem: PPO, A2C, SAC, TD3. According to the simulation results, TD3 shows the best results, outperforming MILP by 5 % in cost savings, and the time to solve the problem is reduced by about a factor of three.

Keywords: reinforcement learning, energy management system, distributed energy system, numerical optimization.

1. Introduction. Due to environmental problems, increased demand for energy, unstable pricing policy for fuel resources and lack of energy capacity, attention has been focused on distributed energy technologies (DRE). Distributed energy resources are generally small-scale sources of energy generation and storage located in close proximity to the place of use of electricity, they can provide an alternative or improvement to the traditional electrical grid. Distributed energy technologies include gas piston, gas turbine and

* This work was supported by St. Petersburg State University (project ID: 94062114).

micro-turbine power plants, heat pumps, steam boilers, renewable energy (solar panels, wind generators), energy storage, fuel cells, co-generation plants, etc. Together, they offer consumers the potential for cost reduction, energy independence and increased energy efficiency.

For the implementation and effective use of RE technologies, a control mechanism is needed that controls the processes occurring in the user's power system. The main goals of energy management are: resource conservation, pollution control and cost savings, provided that users have constant access to the energy they need. The introduction of RE technologies with a control system into the electric power system (EPS) of the user forms the concept of Smart Grid (smart networks), where a potential electric consumer of any level gets the opportunity to interact with the EPS: predict and plan consumption, choose a supplier and influence tariffs. The main attributes of the Smart Grid concept are defined as follows: availability, reliability, flexibility, efficiency, safety, energy storage capacity, stimulation of the activity of the electric consumer, reduction of environmental pressure on the environment.

Renewable energy sources and energy storage systems play a critical role in optimal microgrid planning. Energy storage system can increase the flexibility of a smart grid, while renewable energy provides partial or complete independence from the utility. The researchers were faced with the task of developing an optimization model that minimizes the cost of purchasing electricity by scheduling charging and discharging batteries, as well as exchanging with the energy market, subject to system constraints and achieving energy balance.

The research work has following contribution:

1) designed a piece-wise reward function suitable for energy trading mode and the effective state space of the agent with the important features that most affect the agent's behavior;

2) compared performance of classical optimization method MILP and state of the art deep reinforcement learning method such as PPO, TD3, SAC, A2C in terms of battery energy management. TD3 outperformed MILP in terms of final cost reduction and calculation time.

The reminder of the paper is orginized as follows. Section 2 describes existing optimization methods that can be used for energy management system. Section 3 describes power system, simulation mechanism and data. Section 4 introduce MILP and reinforcement learning based optimization solution for battery energy scheduling. Section 5 presents simulation results and conclusion is made in Section 6.

2. Related work. Research in the field of optimal battery management covers a wide range of scientific areas and tasks. One of these areas focuses on optimal battery management to minimize financial costs or environmental impact. In recent years, several research papers have been published to control energy saving systems using various approaches, where mixed integer linear programming (MILP) is used to model the power system. In [1], software packages or "solvers" are used to solve the problem. Energy management of microgrids using fuzzy logic is discussed in [2]. Intelligent energy management of microgrids using a genetic algorithm is discussed in works [3, 4]. Energy management of hybrid renewable energy generation using limited optimization was proposed in [5]. The expert system and other classical and heuristic microgrid energy management algorithms are discussed in [6, 7].

These studies cannot take into account all the factors and problems that arise in the problems of managing storage elements in hybrid power systems. Designing a controller

based on a model requires precise input data and method parameters to solve the problem. Under the conditions of the task at hand, this can be hampered by the heterogeneous and dynamic nature of electricity consumption and the intermittent nature of renewable energy generation. In addition, many algorithms require large computing power and cannot be adapted to solve real industrial problems. Therefore, a method based on reinforcement learning (RL) is adopted as a solution.

RL is a branch of artificial intelligence that investigate how agent learn policy by interacting with environment. RL has gained huge attention after RL combined successfully with deep learning and achived good performance on Atari. The ability to find optimal in high dimensional search space made RL a strong candidates for nonlinear optimization problems [8]. The RL approach learns optimal strategies through a trial and error mechanism, and does not require a description of the distribution of uncertainties in datasets, as this method adapts and learns to automatically capture and use the set of uncertainties contained in historical data, RL has gained huge attention in many fields such as autonomous driving, robotics [9].

As the most popular RL method used in the energy field, Q-learning is used as a solution to the battery energy scheduling problem, capable of finding the optimal solution given the battery degradation [10]. Q-learning and deep Q-networks, have shown good performance for wind farm control in tracking maximum power points [11], in local energy trading strategies [12], and many other uncertainty problems [13]. Without specifying the exact model and its parameters, the RL algorithm can determine the control strategy by extracting effective characteristics from the data. However, in the context of the task of managing an energy storage system, multiple decision parameters, control variables, and inevitable uncertainties in the data lead to a multidimensional and continuous space of states and actions. This results in the slow convergence rate of the aforementioned RL algorithms that discretize a continuous state space and discard observations after each update, resulting in inefficient data usage and affecting optimization results. To solve this problem, [14] uses a deep deterministic policy gradient (DDPG) to control charge/discharge power. Authors [15] used DDPG to optimise a household solar PV battery system. It does not require special statistical models and discretization of continuous problems [16], and does not require a description of the distribution of uncertainties in data sets, since this method adapts and learns to automatically capture and use the set of uncertainties contained in historical data. However, these scientific papers used fairly simple reward functions that significantly affected convergence when the agent was trained in the environment. In this work, we developed a piece-wise reward function suitable for energy trading mode and the effective state space of the agent with the important features that most affect the agent's behavior. In recent years, an increasing number of reinforcement learning algorithms for continuous environments have emerged that are less well researched in the context of the battery station control problem. The paper [17] provides a benchmarking analysis of deep RL for continuous control, according to the results of which the most effective algorithms based on proximal policy optimization and twin-delayed deep deterministic policy gradient. Thus, we also compare the state-of-the-art RL algorithms for the battery station control problem, to determine the most efficient algorithm in this area.

3. DER control problem statement. This section will provide a description of the energy system, the simulation, and data for the simulation.

3.1. Power system description. The energy system considered in this paper consists of a photovoltaic plant, a battery as an energy storage device, a residential load, in-

verters, and a transformer connecting the microgrid to the local utility network. Inverters convert the direct current (DC) from the battery and PV system into alternating current (AC) for supply to the user's grid. The residential load can be met by using energy from the local photovoltaic system or by purchasing energy from the local utility grid. Surplus energy produced at low energy demand or high production can be stored in a battery and reused during peak demand, or sold to the local utility grid. At time t, the control system requests the necessary information from the database: tariffs, electricity price, predicted values of PV plant generation and load, as well as equipment characteristics. The built-in algorithm should determine $\mathrm{SoC}(t+1)$ — the remaining battery power for the next point in time and transfer the resulting value to the controller. The control element then sends commands to various systems to optimally control the user's power system. The microgrid architecture described is shown in Figure 1.

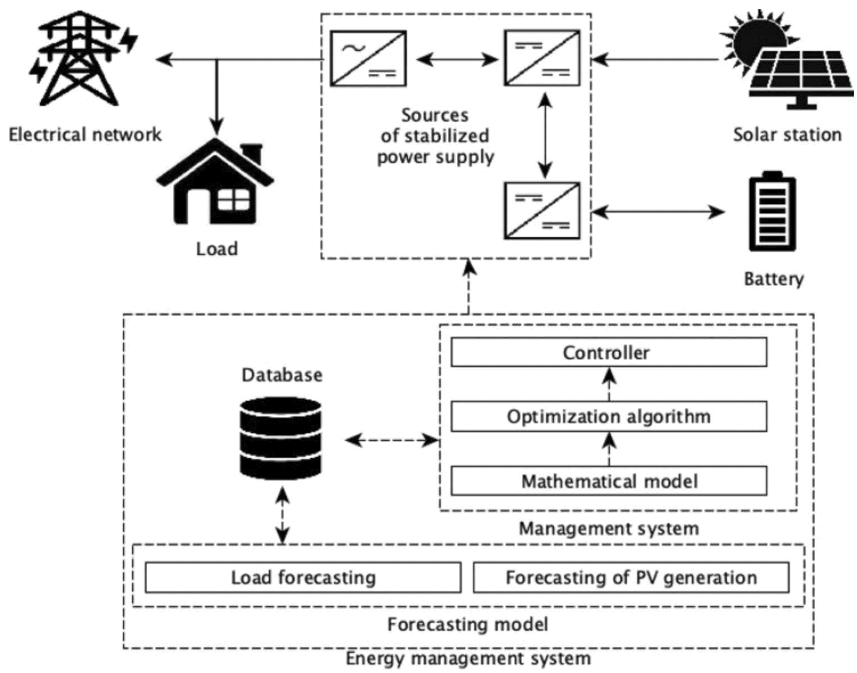

Figure 1. Power system model
The dotted line is the data flow, the straight line is the energy flow.

3.2. Simulation environment. The simulation engine passes data to the battery controller at each time step and queries the SoC for the next time step. Once the battery SoC is established, the energy required to satisfy the energy balance is calculated. If the value obtained is negative, the energy will be purchased from the network in the required amount, if it is positive, then this amount of energy is sold to the network. Then the cost of buying/selling energy is calculated and added to the current amount. In order to obtain the optimization results, also the value of the user's energy purchase costs without the use of storage batteries was calculated. In this case, the load was balanced with the electricity of the main grid and produced by the photovoltaic plant.

3.3. Data description. In this study, we used a data set provided by Schneider Electric [18] — french power engineering company, manufacturer of equipment for power

subcomplexes of industrial enterprises, civil and residential construction facilities, data centers, which has publicly available [19] data that has been proposed to solve energy management problems. Dataset includes several test cases with description of equipment characteristics, forecasted energy production and consumption data (contain a forecasting error, but are necessary for battery station scheduling), and actual energy production and consumption values for previous steps.

The equipment data contains the following information: charging and discharging efficiency — η^{ch}, η^{dch}, maximum charging and discharging power — $P_{ch}^{\max}, P_{dch}^{\max}$, minimum and maximum state of charge of the battery — $\mathrm{SoC}_{\min}$, $\mathrm{SoC}_{\max}$. In this research, 1 test case is selected for the experiment, the corresponding battery characteristic described in Table 1.

Table 1. **Battery characteristic**

Test case	$\mathrm{SoC}_{\max}$	$\mathrm{SoC}_{\min}$	$P_{ch}^{\max}$	$P_{dch}^{\max}$	η^{dch}	η^{ch}
1	300	0	75	−75	0.950	0.950

Data set contains actual values of energy consumption and photovoltaic power output at the previous step and current tariffs for purchase and sale of energy. The dataset also provides predicted time series values of energy consumption, PV power output for the next day in 15 minute increments. The presence of prediction error affects the effectiveness of the model-based control algorithm, which requires accurate input data for optimal control.

4. Optimization solution. This chapter will describe several approaches to controlling storage stations: based on the MILP model, reinforcement learning. And also a case of an energy system without a battery as a baseline for comparison approaches.

4.1. Baseline: power system without battery. To measure optimization performance, the value of the user's costs for the purchase of energy without the use of storage batteries was also calculated. In this case, the load was balanced with electricity from the central grid and produced by a photovoltaic plant.

4.2. Mixed-integer linear programming. The MILP model is formulated as follow according to the Schneider Electric competition problem statement [1, 19]. Note that the effect of battery aging or degradation is not taken into account, the environment has no effect on battery performance.

Objective function:

$$F = \sum_{t=0}^{95} P_{buy}(t) \cdot C_{buy}(t) + \sum_{t=0}^{95} P_{sell}(t) \cdot C_{sell}(t) \to \min, \tag{1}$$

here $P_{buy}(t), P_{sell}(t)$ are represent purchased and sold energy respectively; $C_{buy}(t), C_{sell}(t)$ are tariff for the purchase and sale of energy. The first part of the formula (1) is the total cost for electricity imported from the power grid, the second part refers to the sale of electricity to the grid company during given periods of time. The planning horizon is 96 timestep, which is one day, day of length is chosen since the price policy changes during the days, price policy change almost follows the same pattern during one period. Note that optimization should be taken at every time step.

In the optimization model, the following restrictions are considered for calculating the feasible solution of the cost function.

The energy balance constraint indicates that renewable energy, batteries and grids must meet the grid's electricity demand at every step, as follows:

$$P_{buy}(t) + P_{pv}(t) - P_{dch}(t) = L(t) + P_{ch}(t) - P_{sell}(t) \quad \forall t. \tag{2}$$

In (2) $P_{ch}(t), P_{dch}(t)$ are battery charging and discharging power, $L(t)$ is forecasted load, $P_{pv}(t)$ is predicted power of the photovoltaic plant (these data have a forecasting error). The upper bound of the SoC is limited by rated capacity, and the lower limit is necessary to prevent reducing its lifetime [20]:

$$\mathrm{SoC}_{\min} \leqslant \mathrm{SoC}(t) \leqslant \mathrm{SoC}_{\max}, \tag{3}$$

$$\nu(t) \cdot P_{dch}^{\max} \leqslant P_{dch}(t) \leqslant 0, \tag{4}$$

$$0 \leqslant P_{ch}(t) \leqslant u(t) \cdot P_{ch}^{\max}, \tag{5}$$

$$u(t) + \nu(t) \leqslant 1. \tag{6}$$

In formulas (3)–(6) $P_{ch}^{\max}, P_{dch}^{\max}$ are maximum charging and discharging power of the battery, $\mathrm{SoC}_{\min}$, $\mathrm{SoC}_{\max}$ are minimum and maximum battery charge status. The SoC of battery is expressed as [20]

$$\mathrm{SoC}(t) = \mathrm{SoC}(t-1) + P_{ch}(t-1) \cdot \eta^{ch} + P_{dch}(t-1)/\eta^{dch} \quad \forall t, t \neq 1,$$

where η^{ch}, η^{dch} are battery charging and discharging efficiency; $u(t), \nu(t) - 1$, if the battery is discharged, 0 otherwise:

$$\mathrm{SoC}(0) = \mathrm{SoC}_{\mathrm{In}},$$

$$P_{buy}(t) \geqslant 0,\ P_{dch}(t) \leqslant 0,\ P_{ch}(t) \geqslant 0,\ P_{sell}(t) \leqslant 0 \quad \forall t,$$

$\mathrm{SoC}_{\mathrm{In}}$ is initial battery charge, $\mathrm{SoC}(t)$ is state of charge battery.

4.3. Reinforcement Learning. RL refers to learning how to take actions in different states of an environment to maximize cumulative future reward, where the reward is a form of feedback associated with a task and received in response to an action [21]. RL process is usually modeled as Markov Decision Processes (MDPs). An MDP process is defined using a tuple $< S, A, P, r >$. Parameters S is the state space of the environment, where the state $s \in S$ represents a situation of the environment; A is the set of actions that can be taken by an agent; $r : S \times A \to \mathbb{R}$ is the numerical reward obtained as a function of state and action. The goal of RL is to maximize future discounted return. The value of taking action a_t in s_t is calculated as the expected cumulative reward obtained over all possible trajectories, as follows [21]:

$$Q^{\pi}(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{a \sim \pi(s)} \left[\sum_{i=1}^{\infty} \gamma^i r(s_{t+i}, a_{t+i}) \mid s_t, a_t \right].$$

Here $Q^{\pi}(s_t, a_t)$ is called the Q-value of the state-action pair while an agent follows a policy π.

Advantage Actor Critic (A2C) [22] is a actor critic method, relies on n-step updating. A2C is a synchronous, deterministic implementation that waits for each actor to finish its segment of experience before updating, averaging over all of the actors.

Proximal Policy Optimization (PPO) [23] is policy gradient method for RL which alternate between sampling data through interaction with the environment, and optimizing a "surrogate" objective function using stochastic gradient ascent.

Twin-Delayed Deep Deterministic Policy Gradient (TD3) [24] is a improved version of DDPG, used the idea behind the Double DQN to tackle overestimation bias with the value function.

Soft Actor Critic (SAC) [25] is an off-policy actor-critic deep RL algorithm based on the maximum entropy reinforcement learning framework. In this framework, the actor aims to maximize expected reward while also maximizing entropy. That is, to succeed at the task while acting as randomly as possible.

4.4. Deep reinforcement learning based battery controller. In order to model the battery energy scheduling task as MDP, we define state space, action space and reward function as follow.

State space (δ). State space consists of a time component is S_t, uncontrolled exogenous component is S_x and the controlled part is S_c :

$$
\begin{aligned}
&S = S_t \times S_x \times S_b,\\
&S_t = S_t^d \times S_t^q \times S_t^m \times S_t^h,\\
&S_x = S_x^l \times S_x^{pv} \times S_x^b \times S_x^s \times S_x^{l_0} \times S_x^{pv_0} \times S_x^{b_0} \times S_x^{S_0},
\end{aligned}
$$

where S_t^q is represents a quarter of an hour of the day; S_t^d is day of the week; S_t^m represents month; S_t^h represents hour of the day; S_x^l is vector of two timestamp predicted values of residential load; S_x^{pv} is vector of two timestamp forecast values of the energy generated by a photovoltaic station; S_x^b, S_x^s are vectors of tariff values for the purchase and sale of energy, respectively; $S_x^{l_0}, S_x^{pv_0}, S_x^{b_0}, S_x^{S_0}$ are current values of the residential load, the energy of the photovoltaic plant, the tariffs for the purchase and sale of energy at the previous time step; S_b is represents battery energy level. The state features are normalized only before feed to the deep RL network, in simulation environment the state features are not normalized.

Action Space (A). In any moment of time t the agent can perform one of the following actions $a_t \in [P_{dch}^{\max}; P_{ch}^{\max}]$.

Reward function (R) is designed to encourage agent to save extra energy by charging battery when PV larger than electricity demand, and discharge battery when PV production is not enough to cover electricity demand. The buy price is always bigger than the sell price, thus it's not a wise move to sell energy back to main grid:

$$
R_t = \begin{cases} \arctan(P_t^2 \cdot (\frac{1.5}{\max(P_{buy})^2}) \cdot (-\frac{P_{ch}(t)}{E_{\text{grid}}}) - \text{Penalty} & \text{if } E_{\text{grid}} > 0, \\ \arctan(\frac{1}{\max(0.000001, abs(E_{\text{net}})}) - \text{Penalty} & \text{if } E_{\text{grid}} < 0. \end{cases} \tag{7}
$$

In (7) $P_t = P_{sell}$ if net grid energy is smaller than 0, $P_t = P_{buy}$ if net grid energy is larger than 0, $\max(P_{buy})$ is represents largest historical energy buy price. Penalty is the absolute value of difference of agent proposed battery energy level and battery capacity, range between [0, 1]. Note that arctan function is used to scale reward values. $P_{ch}(t)$ is amount of battery discharged/charged energy; E_{grid} is grid energy without battery, which is the difference of electricity demand and PV; E_{net} is net grid energy, which is sum of battery charged/discharged energy and electricity demand minus amount of PV produced energy.

The first part of the piece-wise reward function is designed to encourage agent to discharge battery energy to cover energy deficit. $-\frac{P_{ch}(t)}{E_{\text{grid}}}$ is designed to encourage agent to discharge energy enough for covering energy deficit, value of $P_{ch}(t)$ is negative when battery discharges. Agent might choose to discharge battery as much as possible to get

higher reward, thus energy buy price and sell price is taken into consideration, prevent battery discharge too much at lower energy buy price period. The second part of the piecewise reward function is designed to save extra energy produced by PV station, the higher the agent saves PV produced energy the higher reward it gets.

5. Simulation results. The simulation is performed on Dell Inc. — Dell G15 5510, with processor Intel(R) Core(TM) i7-10870H CPU 2.20GHz. The Mixed Integer-Linear Programming formulation solved using solver Gurobi [26] with the following parameters: time limit of 15 minutes, relative MILP optimality gap of 0.03, and deterministic concurrent method are used each MILP solution. Models of reinforcement learning were implemented and trained using framework OpenAI Gym [27].

5.1. Model training with reinforcement learning. Each model is represented by a full-connected neural network with 2 hidden layers with 256 neurons on each layer, and activation function used hyperbolic tangent. For training, the discount factor is set to 0.99, and the Adam optimizer with constant learning rate $2 \cdot 10^{-5}$ is used. Other parameters follow the default settings of each algorithms in Gym.

The RL models for controlling the battery station have been trained using state-of-the-art reinforcement learning algorithms: PPO, A2C, SAC, TD3 on a separate training data set.The training process is visualized at Figure 2. TD3 and SAC show better convergence than PPO, A2C, but require more iterations. In the next section, these trained models for each algorithm will be tested on a benchmark dataset.

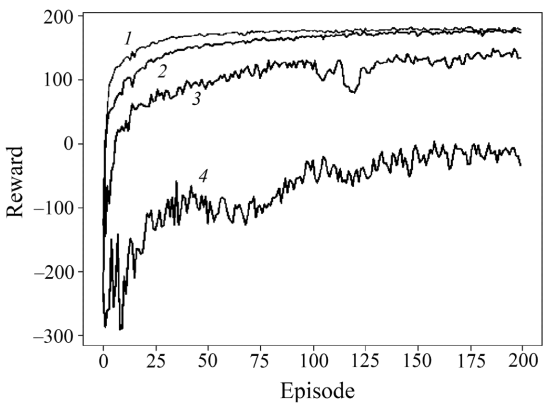

Figure 2. Learning curves for SAC (*1*), TD3 (*2*), PPO (*3*) and A2C (*4*)

5.2. Result compared RL-based and MLIP-based algorithms. To present the results of optimization, we compare the execution time of the algorithm — Time (s), as one of the important indicators in modern energy management, and use a metric — Score, which is equal to the average value of the relative costs of purchasing electricity using a battery, to the cost of buying electricity without a battery:

$$\text{Score} = \frac{\text{Money}_{spent} - \text{Money}_{no\_batt}}{|\text{Money}_{no\_batt}|},$$

where Money_{spent} is the cost with the battery and the scheduling algorithm; Money_{no_batt} is the cost without the battery.

The results comparing PPO, A2C, SAC and TD3 and the deterministic approach are shown in Table 2.

According to experimental results, all models except A2C provide a reduction in user costs comparable to the results of the exact MILP simulation. Moreover, the models trained

Table 2. **Comparison of MILP, PPO, TD3, SAC, A2C**

Method	Money_{spent}	Money_{no_batt}	Score	Time, s
MILP	1162.88	1233.09	–0.0569	25.73
PPO	1170.19	1233.09	–0.0510	7.09
A2C	1208.08	1233.09	–0.020	6.21
SAC	1162.09	1233.09	–0.0575	7.41
TD3	1155.82	1233.09	–0.0626	8.10

with the TD3 algorithm outperform the MILP results by about 5 % on the Score metric. This is a consequence of the fact that RL methods can adapt and detect uncertainties in the historical data during training. In addition, the computation time in reinforcement learning approaches decreases by more than 3.5 times, which indicates that RL methods can be scaled up to solve real production problems.

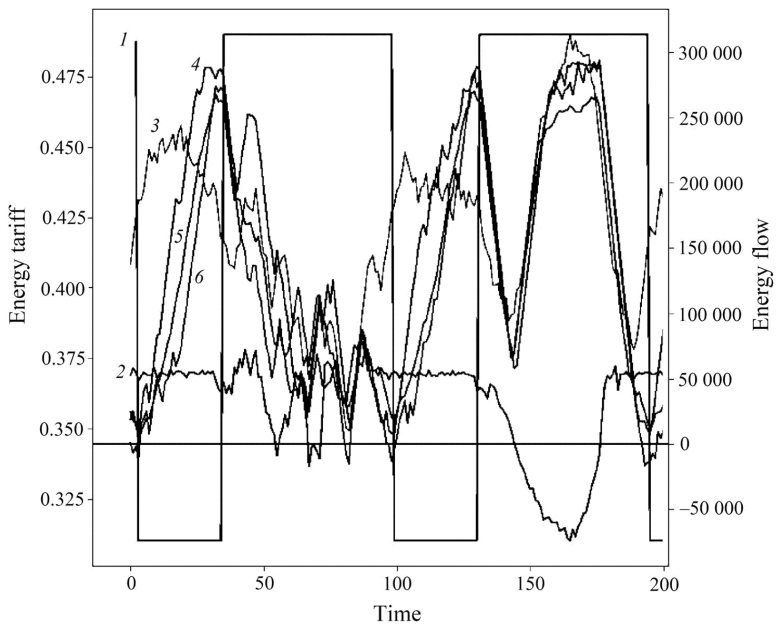

Figure 3. Power consumption and battery charging/discharging profiles for RL-based controller
1 — energy purchase price; *2* — grid energy; *3* — A2C battery controller; *4* — TD3 battery controller; *5* — PPO battery controller; *6* — SAC battery controller.

The simulation results shown at Figure 3 indicate that PPO, A2C and SAC charge and discharge battery smoothly than TD3, and do not always respond to trend changes in the data. Since energy buying price is always larger than energy sell price, it is not a wise move to sell energy back to main grid. The optimal strategy, which all agents have learned successfully, is to charge battery at low energy buying price period regardless of energy deficit, then discharge battery at high energy buying price period to avoid buying energy from the main grid.

6. Conclusion. In the era of smart grids, the need to implement an efficient control component in the users power system is increasing every year. In this research, existing approaches and practices for optimizing power consumption are reviewed. The most

promising trend in the context of the problem at hand is RL, which has the advantage of self-learning and explores optimal strategies through a trial-and-error mechanism in a dynamic environment.

Therefore, we provide a comparison of the classical optimization method MILP and deep RL optimization approach are applied to power supply systems with distributed energy resources, including the generation of renewable energy from a photovoltaic station and an battery energy storage system. A reward function suitable for trading energy with main grid is designed, and proved to be able to drive deep RL agent to learn optimal strategy to outperform MILP solution. There is also a comparison of four state-of-the-art RL algorithms for the battery power plant control problem: PPO, A2C, SAC, TD3. According to the comparison results, the best algorithm for this problem is TD3, which outperforms MILP by 5 % in terms of cost reduction. Moreover, due to the comparatively low computational complexity of the inference, deep learning models are 3.5 times faster than exact methods, which indicates their scalability for solving large-scale industrial problems.

RL is capable of switching charge at maximum power to discharge at maximum power instantly, which is bad for energy storage equipment. The future research direction will be control battery smoothly, develop method for more equipment friendly control method. Also, as further research to increase the energy efficiency of the modeled system will be considered distributed energy technologies, such as cogeneration systems. This requires additional research and a more detailed study of the process of co-generation of electricity and heat, as well as a larger set of historical data.

References

1. Markelova A., Petrosian O., Laistseva M. Microgrid management with energy storage system and renewable energy generation. *Control Processes and Stability*, 2021, vol. 8, no. 1, pp. 430–434.

2. Hatziargyriou N. D. Special issue on microgrids and energy management. *European transactions on electrical power*, 2011, vol. 21, no. 2, pp. 1139–1141.

3. Reddy P. P., Veloso M. M. Strategy learning for autonomous agents in smart grid markets. *Twenty-Second International Joint Conference on Artificial Intelligence*. Barcelona, 2011, pp. 1446–1451.

4. Chen C., Duan S., Cai T., Liu B., Hu G. Smart energy management system for optimal microgrid economic operation. *IET Renewable Power Generation*, 2011, vol. 5, no. 3, pp. 258–267.

5. Mohamed F. A., Koivo H. N. System modelling and online optimal management of microgrid with battery storage. *6^{th} International Conference on renewable energies and power quality (ICREPQ'07)*. Sevilla, 2007, pp. 26–28.

6. Colson C., Nehrir M., Pourmousavi S. Towards real-time microgrid power management using computational intelligence methods. *IEEE PES general meeting*, IEEE, 2010, pp. 1–8.

7. Abdilahi A. M., Mustafa M., Aliyu G., Usman J. Autonomous Integrated Microgrid (AIMG) system: Review of Potential. *International Journal of Education and Research*, 2014, vol. 2, no. 1, pp. 1–18.

8. Mnih V., Kavukcuoglu K., Silver D., Graves A., Antonog I., Wierstra D., Riedmiller M. Playing atari with deep reinforcement learning. *arXiv preprint*, 2013, arXiv: 1312.5602.

9. Perera A., Kamalaruban P. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 2021, vol. 137, pp. 1–22.

10. Muriithi G., Chowdhury S. Optimal energy management of a grid-tied solar pv-battery microgrid: A reinforcement learning approach. *Energies*, 2021, vol. 14, no. 9, pp. 1–24.

11. Wei C., Zhang Z., Qiao W., Qu L. Reinforcement-learning-based intelligent maximum power point tracking control for wind energy conversion systems. *IEEE Transactions on Industrial Electronics*, 2015, vol. 62, no. 10, pp. 6360–6370.

12. Xi L., Yu L., Fu Y., Huang Y. Automatic generation control based on deep reinforcement learning with exploration awareness. *Proceedings of the CSEE*, 2019, vol. 39, no. 14, pp. 4150–4162.

13. Wang B., Zhou M., Xin B., Zhao X., Watada J. Analysis of operation cost and wind curtailment using multi-objective unit commitment with battery energy storage. *Energy*, 2019, vol. 178, pp. 101–114.

14. Gao Y., Yang J., Yang M., Li Z. Deep reinforcement learning based optimal schedule for a battery

swapping station considering uncertainties. *IEEE Transactions on Industry Applications*, 2020, vol. 56, no. 5, pp. 5775–5784.

15. Kell A. J., McGough A. S., Forshaw M. Optimizing a domestic battery and solar photovoltaic system with deep reinforcement learning. *arXiv preprint*, 2021, arXiv: 2109.05024.

16. Wu Y., Tan H., Peng J., Zhang H., He H. Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus. *Applied Energy*, 2019, vol. 247, pp. 454–466.

17. Duan Y., Chen X., Houthooft R., Schulman J., Abbeel P. Benchmarking deep reinforcement learning for continuous control. *International Conference on Machine Learning*. New York, 2016, pp. 1329–1338.

18. *Official website Schneider Electric*. Available at: https://www.se.com/ww/en/ (accessed: March 1, 2022).

19. Power Laws: Optimizing Demand-side Strategies. Available at: https: //www.drivendata.org/competitions/53/optimize-photovoltaic-battery/ (accessed: March 1, 2022).

20. Sedighizadeh M., Esmaili M., Mohammadkhani N. 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. *Journal of Cleaner Production*, 2018, vol. 195, pp. 301–317.

21. Sutton R. S., Barto A. G. *Reinforcement learning: An introduction*. London, MIT Press, 2018, 590 p.

22. Babaeizadeh M., Frosio I., Tyree S., Clemons J., Kautz J. Reinforcement learning through asynchronous advantage actor-critic on a GPU. *arXiv preprint*, 2016, arXiv: 1611.06256.

23. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithms. *arXiv preprint*, 2017, arXiv: 1707.06347.

24. Fujimoto S., Hoof H., Meger D. Addressing function approximation error in actor-critic methods. *International Conference on Machine Learning*, PMLR, 2018, pp. 1587–1596.

25. Haarnoja T., Zhou A., Abbeel P., Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International Conference on Machine Learning*, PMLR, 2018, pp. 1861–1870.

26. Mittelmann H. Latest benchmark results. *Proceedings of the INFORMS Annual Conference*. Phoenix, 2018, pp. 4–7.

27. Brockman G., Cheung V., Pettersson L., Schneider J., Schulman J., Tang J., Zaremba W. OpenAI Gym. *arXiv preprint*, 2016, arXiv: 1606.01540.

Received: May 26, 2023.
Accepted: June 8, 2023.

Authors' information:

Anastasia Yu. Zhadan — Student; e-mail anastasiya_markel@mail.ru

Haitao Wu — Student; a89173627785@gmail.com

Pavel S. Kudin — Student; accpavel1@gmail.com

Yuyi Zhang — Student; Lesliezhang0825@gmail.com

Ovanes L. Petrosian — Dr. Sci. in Physics and Mathematics, Professor; petrosian.ovanes@yandex.ru

Управление микроэнергосистемой с возобновляемыми источниками энергии на основе глубокого обучения с подкреплением и подходов численной оптимизации*

А. Ю. Жадан, Х. Ву, П. С. Кудин, Ю. Чжан, О. Л. Петросян

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

* Работа выполнена при финансовой поддержке Санкт-Петербургского государственного университета (ID проекта: 94062114).

Для цитирования: *Zhadan A. U., Wu H., Kudin P. S., Zhang Y., Petrosian O. L.* Microgrid control for renewable energy sources // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2023. Т. 19. Вып. 3. С. 391–402. https://doi.org/10.21638/11701/spbu10.2023.307

Оптимальное планирование работы аккумуляторной системы хранения энергии играет важную роль в распределенной энергетической системе. Как метод, основанный на данных, глубокое обучение с подкреплением не требует наличия системных знаний о динамической системе, позволяя найти оптимальное решение для нелинейной задачи оптимизации. В данном исследовании финансовые затраты на потребление энергии снижены за счет планирования энергии аккумуляторов с использованием метода глубокого обучения с подкреплением (RL). Обучение с подкреплением может адаптироваться к изменениям параметров оборудования и шумам в данных, в то время как смешанно-целочисленное линейное программирование (MILP) требует высокой точности прогнозирования выработки и спроса на электроэнергию, точных параметров оборудования для достижения хорошей производительности, а также больших вычислительных затрат для крупномасштабных промышленных приложений. Исходя из этого, можно предположить, что решение на основе глубокого RL способно превзойти классическую детерминированную модель оптимизации MILP. Сравниваются четыре современных RL-алгоритма для задачи управления аккумуляторной электростанцией: PPO, A2C, SAC, TD3. Согласно результатам моделирования, TD3 показывает наилучшие результаты, превосходя MILP на 5 % по экономии средств, а время решения задачи сокращается примерно в 3 раза.

Ключевые слова: обучение с подкреплением, система управления энергией, распределенная энергетическая система, численная оптимизация.

Контактная информация:

Жадан Анастасия Юрьевна — студент; anastasiya_markel@mail.ru

Ву Хайтао — студент; a89173627785@gmail.com

Кудин Павел Сергеевич — студент; accpavel1@gmail.com

Чжан Юйи — студент; Lesliezhang0825@gmail.com

Петросян Ованес Леонович — д-р физ.-мат. наук, проф.; petrosian.ovanes@yandex.ru