Performance Comparison of Deep RL Algorithms for Energy Systems Optimal Scheduling

Shengren, Hou; Salazar , Edgar Mauricio; Vergara , Pedro P.; Palensky, Peter

DOI
10.1109/ISGT-Europe54678.2022.9960642

Publication date
2022

Document Version
Final published version

Published in
Proceedings of 2022 IEEE PES Innovative Smart Grid Technologies Conference Europe, ISGT-Europe 2022

Citation (APA)
Shengren, H., Salazar , E. M., Vergara , P. P., & Palensky, P. (2022). Performance Comparison of Deep RL Algorithms for Energy Systems Optimal Scheduling. In *Proceedings of 2022 IEEE PES Innovative Smart Grid Technologies Conference Europe, ISGT-Europe 2022* (pp. 1-6). (IEEE PES Innovative Smart Grid Technologies Conference Europe; Vol. 2022-October). IEEE. https://doi.org/10.1109/ISGT-Europe54678.2022.9960642

Performance Comparison of Deep RL Algorithms for Energy Systems Optimal Scheduling

Hou Shengren[1], Edgar Mauricio Salazar[2], Pedro P. Vergara[1], Peter Palensky[1]
[1]Intelligent Electrical Power Grids, Delft University of Technology, The Netherlands
[2]Electrical Energy Systems (EES) Group, Eindhoven University of Technology, The Netherlands
emails: {H.Shengren, P.P.VergaraBarrios, P.Palensky}@tudelft.nl, E.M.Salazar.Duque@tue.nl

2022 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe) | 978-1-6654-8032-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISGT-Europe54678.2022.9960642

***Abstract*—Taking advantage of their data-driven and model-free features, Deep Reinforcement Learning (DRL) algorithms have the potential to deal with the increasing level of uncertainty due to the introduction of renewable-based generation. To deal simultaneously with the energy systems' operational cost and technical constraints (e.g, generation-demand power balance) DRL algorithms must consider a trade-off when designing the reward function. This trade-off introduces extra hyperparameters that impact the DRL algorithms' performance and capability of providing feasible solutions. In this paper, a performance comparison of different DRL algorithms, including DDPG, TD3, SAC, and PPO, are presented. We aim to provide a fair comparison of these DRL algorithms for energy systems optimal scheduling problems. Results show DRL algorithms' capability of providing in real-time good-quality solutions, even in unseen operational scenarios, when compared with a mathematical programming model of the energy system optimal scheduling problem. Nevertheless, in the case of large peak consumption, these algorithms failed to provide feasible solutions, which can impede their practical implementation.**

***Index Terms*—Energy management, Machine learning, Deep learning, Reinforcement learning,**

NOTATION

Sets:

$\mathcal{G}$	Set of (DGs) distributed generators
$\mathcal{B}$	Set of ESSs
$\mathcal{L}$	Set of Loads
$\mathcal{V}$	Set of PVs
$\mathcal{S}$	Set of states of the RL algorithms
$\mathcal{A}$	Set of actions
$\mathcal{R}$	Set of rewards
$\mathcal{T}$	Set of time steps

Indexes:

i	DG unit $i \in \mathcal{G}$
j	ESS $j \in \mathcal{B}$
m	PV unit $m \in \mathcal{V}$
k	Load demand $m \in \mathcal{L}$
t	Time-step $t \in \mathcal{T}$

Parameters:

$a_i\ b_i\ c_i$	Quadratic, linear and constant parameters associated to the i-th DG operation cost
Δt	Length for the discretization of the operational time
λ	Discount factor
$\overline{P}_t^G\ \underline{P}_i^G$	Maximum and minimum generation limit of the DG units
$RU_i\ RD_i$	Ramping up and ramping down ability of the DG units
$\overline{P}_j^B\ \underline{P}_j^B$	Maximum and minimum charging/discharging limit of the ESSs
$\overline{E}_j^B\ \underline{E}_j^B$	Maximum and minimum level of SOC of the ESSs
$\overline{P}^C$	Maximum main grid export/import limit
β	Electricity sell coefficient
η_B	Energy exchange efficiency for ESSs
$\sigma_1\ \sigma_2$	Reward re-scale and constrain penalty coefficients
ρ_t	Electricity price for time slot t
$P_{m,t}^V$	Active power of PV systems
$P_{k,t}^L$	Active power demand

Continuous Variables:

$P_{i,t}^G$	Active power output of DG units
$P_{j,t}^B$	Active power discharge/charge of ESSs
$SOC_{j,t}^B$	State of charge for ESSs
P_t^N	Active power exported/imported of main grid
ΔP_t	Active power unbalance

Binary Variables:

$u_{i,t}$	Operational state of the DG units

I. INTRODUCTION

The massive integration of renewable-based resources inherently increases the complexity of the energy systems' scheduling problem [1]. Existing scheduling approaches are not adequate to deal with a large number of distributed generation (DG) units and the increasing level of uncertainty. Innovative approaches are urgently needed, capable of providing in real-time good-quality solutions while still enforcing operational and technical constraints. In the technical literature, two main approaches are available to deal with the energy system's uncertainty, namely *model-based* and *model-free* approaches [2]. In model-based approaches, uncertainty is considered either by using a probability distribution function or by leveraging a set of representative scenarios, constructing a complex mathematical formulation considering the system's operational constraints. This formulation leads to a stochastic multi-stage

mathematical programming model that can be solved using commercial mixed-integer nonlinear programming (MINLP) solvers [3], [4], such as CPLEX. Nevertheless, although capable of providing good quality solutions, existing model-based approaches are not adequate for real-time operation due to the large computational time required.

To overcome this, model-free approaches have been introduced as an alternative solution. The most promising model-free approach is based on the use of reinforcement learning (RL) [5]. In this, by modeling the decision-making problem as a Markov Decision Process (MDP), RL algorithms can learn the system's dynamics by interaction, providing good-quality solutions guided by a reward value used as a performance indicator [6]. The main advantage of RL-based algorithms is that, if properly designed and trained, they can capture energy systems' uncertainty by leveraging historical data, providing good-quality solutions in real-time, and avoiding any computational burden during operation.

Recently, Deep reinforcement learning (DRL), using Deep Neural Networks (DNN) as function approximators, has shown good performance on solving different energy-related scheduling problems with continuous actions, including home appliances [7] and microgrids scheduling [8]. Unfortunately, these works rely on value-based RL algorithms, which do not scale well. Instead, policy-based DRL algorithms, by integrating an actor-critic structure, have shown outstanding performance on complex MDP tasks such as robot control and video games [9]. The current DRL state-of-the-art includes algorithms such as Proximal Policy Optimization (PPO) [10], Soft Actor-Critic (SAC), Policy optimization [11], Deep Deterministic Policy Gradient (DDPG) [12] and derivation (TD3) [13].

Unlike general MDP tasks, the energy system optimal scheduling problem, formulated as an MDP, must enforce a rigorous set of operational constraints to ensure a reliable operation. For instance, the power balance constraint, modeled as an equality constraint, must always be met during real-time operation. Enforcing equality constraints in DRL algorithms is currently an unresolved challenge. Several works indirectly enforced the power balance constraint, for example, by defining a specific DG unit as an infinite source of power [14] or by adding a penalty term to the reward function [15], [16]. Nevertheless, this introduces extra hyperparameters that impact the DRL algorithms' performance and capability of providing feasible solutions. Given the challenge mentioned above, in this paper, we aim to provide a fair performance comparison of DRL algorithms for energy systems optimal scheduling problems. To do this, we compared DDPG, TD3, SAC, and PPO algorithms' performance and performed a sensibility analysis of hyperparameters on the algorithms' capabilities of providing good-quality solutions, even in unseen operational scenarios.

II. Mathematical Programming Formulation

The energy systems optimal scheduling problem can be modeled as the the MINLP model given by (1)-(12). The objective function in (1) aims at minimizing the operational cost for the whole time horizon $\mathcal{T}$, which comprises the operational cost of the DG units, as presented in (2) and the cost of buying/selling electricity from/to the main grid, as in (3). Given the output power of DG units $P^G_{i,t}$, the operational cost can be estimated by using a quadratic model as in (2). The transaction cost between the energy system and the main grid is settled according Time-of-Use (TOU) prices, in which it is assumed that selling prices are lower than the purchasing prices. In (3), ρ_t is the TOU price at time slot t, while P^N_t refers to the exported/imported power transaction to/from the main grid.

$$\min \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{G}}(C^G_{i,t}+C^E_t)\Delta t \tag{1}$$

$$C^G_{i,t} = a_i\cdot\left(P^G_{i,t}\right)^2 + b_i\cdot P^G_{i,t} + c_i, \quad i\in\mathcal{G}. \tag{2}$$

$$C^E_t = \begin{cases}\rho_t P^N_t & P^N_t > 0,\\ \beta\rho_t P^N_t & P^N_t < 0.\end{cases} \tag{3}$$

Subject to:

$$\sum_{i\in\mathcal{G}} P^G_{i,t} + \sum_{m\in\mathcal{V}} P^V_{m,t} + P^N_t + \sum_{j\in\mathcal{B}} P^B_{j,t} = \sum_{k\in\mathcal{L}} P^L_{k,t}, \forall t\in\mathcal{T} \tag{4}$$

$$\underline{P}^G_i\cdot u_{i,t} \le P^G_{i,t} \le \overline{P}^G_i\cdot u_{i,t} \qquad \forall i\in\mathcal{G},\forall t\in\mathcal{T} \tag{5}$$

$$P^G_{i,t} - P^G_{i,t-1} \le RU_i \qquad \forall i\in\mathcal{G},\forall t\in\mathcal{T} \tag{6}$$

$$P^G_{i,t} - P^G_{i,t+1} \le RD_i \qquad \forall i\in\mathcal{G},\forall t\in\mathcal{T} \tag{7}$$

$$-\underline{P}^B_j \le P^B_{j,t} \le \overline{P}^B_j \qquad \forall j\in\mathcal{B},\forall t\in\mathcal{T} \tag{8}$$

$$SOC^B_{j,t} = SOC_{j,t-1} + \eta_B P^B_{j,t}\Delta t \qquad \forall j\in\mathcal{B},\forall t\in\mathcal{T} \tag{9}$$

$$\underline{E}^B_j \le SOC^B_{j,t} \le \overline{E}^B_j \qquad \forall j\in\mathcal{B},\forall t\in\mathcal{T} \tag{10}$$

$$-\overline{P}^C \le P^N_t \le \overline{P}^C \qquad \forall t\in\mathcal{T} \tag{11}$$

$$u_{i,t}\in\{0,1\} \qquad \forall i\in\mathcal{G},\forall t\in\mathcal{T} \tag{12}$$

Expression (4) defines the power balance constrain. A binary variable is used to model the DG unit's commitment status, i.e., $u_{i,t}=1$ represent that the i-th DG unit is operating. Expression (5) defines the DG units generation power limits while (6) and (7) enforce the DG unit's ramping up and down constraints, respectively. Energy storage systems (ESSs) are modeled using (8)-(10). In this model, the operation cost of ESSs is not considered, while ESSs are allowed to schedule their discharge and charge power in advance. Expression (8) defines the charging and discharging power limits, while expression (9) models the state of charge (SOC) as a function of the charging and discharging power. Expression in (10) limits the energy stored in the ESSs, avoiding the impacts caused by over-charging and over-discharging. Finally, main grid export/import power limit is modeled by the expression in (11).

III. MDP Formulation

The above-presented problem can be formulated as a MDP, which can be represented as a 5-tuple $(\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\lambda)$, where $\mathcal{S}$ represents the set of system states, $\mathcal{A}$ the set of actions, $\mathcal{P}$ the state transition probability, $\mathcal{R}$ the reward function, and

λ the discount factor. In this formulation, the energy system operator can be modeled as an RL agent. The state information provides an important basis for the operator to dispatch units. We define a state as $s_t = (P_t^V, P_t^L, P_{t-1}^G, SOC_t), \quad s_t \in \mathcal{S}$, while the actions, defining the scheduling of the DG units and the ESSs, as $a_t = (P_{i,t}^G, P_t^B), \quad a_t \in \mathcal{A}$. Notice that the RL agent does not directly control the transaction between the energy system and the main grid. Instead, after any action is executed, power is exported/imported from the main grid to maintain the power balance. Nevertheless, a maximum power capacity constraint exits and must be enforced i.e., (11).

Given the state s_t and action a_t at time step t, the energy system transit to the next state s_{t+1} defined by

$$\mathcal{P}_{ss'}^a = \Pr\left\{s_{t+1} = s' \mid s_t = s, a_t = a\right\} \tag{13}$$

where $\mathcal{P}_{ss'}^a$ corresponds to the transition probability, which models the energy system's dynamics and uncertainty involved. In model-based algorithms, the uncertainty is predicted by a determined value or sampling from a prior probability distribution. Instead, DRL is a model-free approach, capable of learning the uncertainty and dynamics from historical data and interactions.

The reward r_t is given by the environment as an indicator to guide update direction of the policy. In the energy systems optimal scheduling problem, the reward function should guide the RL agent to take actions that minimize the operational cost while enforcing the power balance constraint. This can be done by using the next reward function

$$r_t\left(s_t, a_t\right) = -\sigma_1 \left[\sum_{i \in \mathcal{G}} C_{i,t}^G + C_t^E\right] - \sigma_2 \Delta P_t, \tag{14}$$

in which ΔP corresponds to the power unbalance at time-step t and is defined as,

$$\Delta P_t = |\sum_{i \in \mathcal{G}} P_{i,t}^G + \sum_{m \in \mathcal{V}} P_{m,t}^V + P_t^N + \sum_{j \in \mathcal{B}} P_{j,t}^B - \sum_{k \in \mathcal{L}} P_{k,t}^L| \tag{15}$$

while σ_2 is used to control the trade-off between the cost minimization and the penalty incurred in case of power unbalance. The goal of the RL algorithms is to find an optimal scheduling policy π^*(stochastic or deterministic) to maximize the total expected discounted rewards for the formulated MDP

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{(s_t, a_t) \sim \rho_\pi} \left[\sum_{t \in \mathcal{T}} R\left(s_t, a_t\right)\right]. \tag{16}$$

IV. DRL Policy-Based Algorithms

The formulated MDP have a continuous state and action spaces, which can be hard to solve by using classical RL algorithms, such as Q learning, due to their poor scalability features [6]. By leveraging the generalization and fitting capability of DNNs, DRL algorithms have shown good performance when dealing with this challenge. In valued-based DRL algorithms, the action-state function Q is iteratively updated to indirectly define a deterministic policy, for which the foundation is Bellman optimality equation:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}\left(s' \mid s, a\right) \max_{a' \in \mathcal{A}} Q^*\left(s', a'\right) \tag{17}$$

If the optimal value function $Q^*(s,a)$ is estimated, then the optimal policy can be derived as $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a)$. Value-based DRL algorithms use DNNs approximate the Q-function, dealing with continuous state spaces. However, for continuous action MDP problems, it requires a full scan of the action space when executing policy improvement i.e., $\arg\max_{a \in \mathcal{A}} Q^\pi(s,a)$, leading to a dimensionality problem. Instead, policy-based DRL algorithms directly search for the optimal policy. The policy is usually modeled with a parametric function, denoted as $\pi_\theta(s|a)$. Based on the policy gradient theorem, the policy gradient is expressed as $\nabla_\theta J(\theta) = \mathbb{E}_\pi\left[Q^\pi(s,a) \nabla_\theta \ln \pi_\theta(a \mid s)\right]$, where $J(\theta)$ is the reward function. Using gradient ascent, θ can be updated towards the direction suggested by $\nabla_\theta J(\theta)$ to find the policy π_θ that lead the highest total discounted rewards.

In this paper, we will compare the performance of state-of-the-art DRL algorithms, including DDPG, TD3, SAC, PPO when solving the MDP described in Sec. III, using a penalty to enforce the power balance constrain, as showed in (14). DDPG and TD3 are off-policy, deterministic algorithms, while PPO and SAC are stochastic algorithms. In general, DRL algorithms interact with the environment to collect a reward. This reward is then used to update a critic DNNs parameters based on the temporal difference (TD) algorithm. Then, the critic network is used to update actor DNNs parameters based on policy gradient theory. A more detailed explanation of policy-based algorithms can be found in [5].

V. Experimental results

A. Experimental Setting

Three DG units, using the parameters shown in Table I, are defined. For the ESSs, the charging and discharging limits are set as 100 kW, the nominal capacity as 500 kW, while the energy efficiency is set as $\eta_B = 0.9$. We assume that the grid's maximum export/import limit is defined as 100 kW. To encourage the use of renewable energies, we set selling prices as half of the current electricity prices, i.e., $\beta = 0.5$. One-hour resolution data profiles including solar generation, demand consumption, and electricity prices for a period of one year are used. The original data is divided into training and testing sets. The training set contains the first three weeks of each month, while the testing sets contain the remaining data. This strategy allows the DRL algorithm to consider seasonal changes in the PV generation and demand consumption. During training, the EESs' initial SOC were randomly set. All algorithms were implemented in Python using PyTorch and trained for 1000 episodes. Unless otherwise mentioned, default settings were used. The implemented DRL algorithms and the environment are openly available at [17]. Hyperparameters σ_1 and σ_2 were defined as 0.01 and 50, respectively, as default values. Each experiment is run with five random seeds to eliminate randomness from the code implementation part during the simulation process. To evaluate

Table I
DG UNITS INFORMATION

Units	a[\$/kW^2]	b[\$/kW]	c[\$]	$\underline{P}^G$[kW]	$\overline{P}^G$[kW]	RU[kW]	RD[kW]
DG_1	0.0034	3	30	10	150	100	100
DG_2	0.001	10	40	50	375	100	100
DG_3	0.001	15	70	100	500	200	200

the DRL algorithms' performance, the total operational cost, as in (1), and the power unbalance as in (15), are used as metrics. To validate and compare the performance of the tested DRL algorithms, we have also solved the MIQP formulation presented in Sec. II using Pyomo [18].

B. Performance on the Training Set

Fig. 1 shows the average reward, losses, operational cost, and power unbalance for all the DRL algorithms during the training process. Reward increased rapidly after training for 100 episodes, while loss decreased to a small value after 200 episodes. This trend is due to the fact that the parameters of all algorithms are initialized randomly, leading to random actions. As a result, a high power unbalance is observed. As the training continues, all DRL agents learn how to meet the power unbalance constraint due to the penalty term used in the reward function. The comparison of the four algorithms shows that the PPO and SAC algorithms are more stable when compared with the DDPG and TD3 algorithms. Not only do PPO and SAC algorithms have a higher reward after training, but they also have a lower variance. In the end, all algorithms converged to similar values. However, DDPG and TD3 algorithms showed a significant performance decrease after training for 800 episodes. For PPO and SAC algorithms, the power unbalance is mitigated notably after training for 100 episodes and nearly disappeared at 1000 episodes, while DDPG and TD3 algorithms show that power unbalances decrease notably after 200 episodes and then remain after training 1000 episodes, at around 500-700 kW.

Fig. 2(a) shows the cumulative operational cost for 10 days in the training set. All tested DRL algorithms have a similar performance compared to MIQP. The solution provided by the MIQP formulation is treated as the optimal solution. Notice that some of the DRL algorithms have outperformed the solution provided by the MIQP formulation in some days in terms of the operational cost. Nevertheless, this is because they did not strictly meet the power balance constraint. Fig. 2(b) shows that the cumulative power unbalances of 10 days is less than 1200 kW, accounting for no more than 5% of the total demand. Notice that the PPO and the TD3 algorithms maintained the lowest and largest power unbalance, respectively, among all algorithms.

C. Performance on the Test Set

Table II and Table III shows the operational cost and the power unbalance, respectively, with 95% confidence intervals, tested on 10 days on the test set. As can be seen, DRL algorithms show similar performance compared with the optimal solutions provided by the MIQP formulation. Among these, the PPO and TD3 algorithms over-perform the DDPG and SAC algorithms regarding the operational cost. Notice that all DRL algorithms were able to operate with a low power unbalance on most of the days from the test set. Nevertheless, all these algorithms failed to maintain the power unbalance in the last test day due to the significantly high demand consumption around peak time. This result shows that in the case of extreme events (e.g., high demand consumption, low renewable generation, etc.), perhaps not captured on the trained data, DRL algorithms fail to provide a feasible solution. These results are important as larger power unbalances can lead to an outage, especially if an export limit is defined with the main grid.

Fig. 3 show the ESSs' SOC and the output power of the DG units, ESSs, and the export/import power from the grid defined using the PPO algorithm and the optimal solution provided by the MIQP formulation. As can be seen, when the electricity price is high and the net power is low, the PPO algorithm dispatches the ESSs in charging mode, while around peak-time, the ESSs is dispatched in discharging mode. A similar operational schedule is defined by the optimal solution provided by the MIQP formulation. As can be seen in Fig. 3, at 12h, the DG unit 1 and the ESSs operating in discharge mode play the most critical role in meeting all the demand. A different operational schedule was defined by the PPO algorithms for the same time step, in which the DG unit 2 and the main grid supply all the demand consumption. In this regard, as the schedule defined by the PPO algorithm did not prioritize the use of the ESSs, a higher operational cost was observed.

D. Sensitivity Analysis

Enforcing technical constraints using a penalty term in the reward function is one of the most common approaches used to provide feasible solutions. As previously mentioned, this approach introduces extra hyperparameters, in this case, to balance the numeric relationship between the operational cost and the power balance constraint. Additionally, as the grid export/import power limit provide flexibility for the actions taken by DRL algorithms, this can significantly impact the feasibility of the provided solutions. In this section, a sensitivity analysis of the penalty coefficient, σ_2, as used in expression (14), is presented.

Fig. 4 shows the convergence performance of all the tested DRL algorithms for different values of the coefficient σ_2 considering $\overline{P}^C = 100$ kW. As can be seen, the operational cost increased significantly when increasing σ_2 from 20 to 50 and then reduced after increasing it from 50 to 100. On the other hand, as expected, increasing σ_2 notably reduced the power unbalance. In this case, if σ_2 is set as 100, the PPO and SAC algorithms were able to almost completely mitigate the power unbalance. Nevertheless, the DDPG and TD3 algorithms were able to minimize the power unbalance for all values of σ_2, which allows concluding that such algorithms are less sensitive to this hyperparameter when compared with the PPO and SAC algorithms.

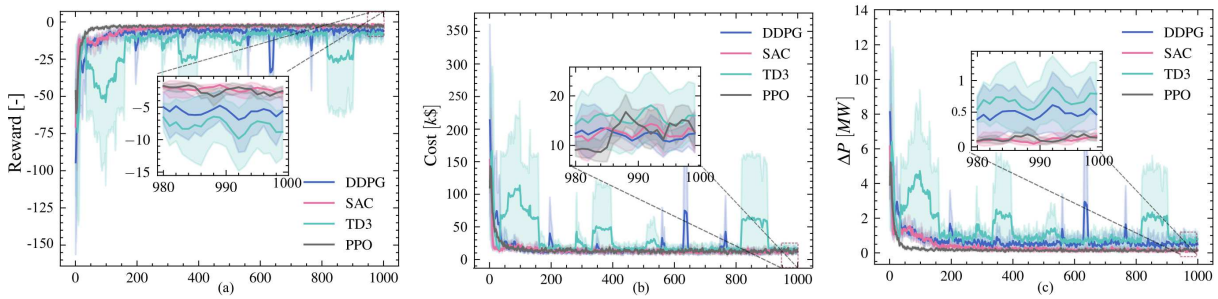


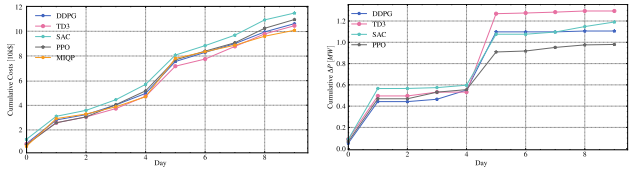
Figure 1. Average (a) reward, (b) operational cost, and (c) power unbalance, during 1000 episode training for all the tested DRL algorithms.

Table II
MEAN AND 95% CONFIDENCE BOUNDS FOR COST[$]

Test day	DDPG	TD3	SAC	PPO	MIQP
1	21142 (20670, 21613)	19810 (19332, 20287)	21907 (21445, 22368)	**17637 (14385, 20888)**	18145
2	11364 (10784, 11943)	**9824 (9306, 10342)**	11084 (8246, 13921)	10833 (9714, 11924)	8358
3	9539 (9162, 9915)	**8868 (8442, 9293)**	10362 (8932, 11792)	10214 (8980, 11446)	7417
4	3292 (1886, 4697)	5120 (4742, 5496)	2502 (1469, 3534)	**2478 (1947, 3009)**	1046
5	23128 (21024, 25232)	23038 (22247, 23829)	20146 (17179, 23112)	**19045 (17399, 20690)**	14971
6	**11266 (10774, 11758)**	15631 (15169, 16092)	13345 (12197, 14491)	12844 (11018, 14669)	8085
7	4046 (2608, 5484)	4901 (4635, 5166)	**2766 (1636, 3894)**	3076 (2254, 3839)	1652
8	11219 (10345, 12092)	**9761 (9424, 10097)**	12019 (8803, 15233)	10778 (8762, 12792)	7126
9	7349 (5838, 8860)	7662 (7190, 8134)	7943 (6017, 9869)	**7250 (5881, 8619)**	5053
10	26285 (24625, 27944)	24050 (23857, 24242)	23303 (22167, 24439)	**22614 (20087, 25141)**	23452

Table III
MEAN AND 95% CONFIDENCE BOUNDS FOR UNBALANCE POWER [KW]

Test day	DDPG	TD3	SAC	PPO
1	189.89 (189.65, 190.14)	246.14 (245.52, 246.76)	225.56 (224.08, 227.04)	**171.27 (29.19, 313.35)**
2	**5.45 (0, 23.46)**	12.63 (12.63, 12.63)	8.30 (0, 39.89)	27.94 (0, 66.88)
3	74.16 (34.41, 113.92)	49.43 (49.43, 49.43)	**25.73 (0, 51.79)**	36.12 (0, 110.21)
4	2.27 (0, 12.64)	37.42 (31.28, 43.57)	**0.01 (0, 0.03)**	36.70 (0.75, 72.65)
5	**10.85 (8.31, 13.39)**	73.40 (73.18, 73.62)	100.75 (70.38, 131.11)	67.81 (0, 185.54)
6	32.34 (32.34, 32.34)	**0.67 (0, 3.59)**	50.05 (35.37, 64.73)	20.52 (0, 84.23)
7	0.07 (0, 0.48)	**0.00 (0, 0)**	37.72 (37.72, 37.72)	3.38 (0, 13.12)
8	64.61 (33.82, 95.41)	107.24 (107.24, 107.24)	32.38 (22.84, 41.91)	66.00 (14.92, 117.08)
9	33.88 (10.52, 57.24)	19.81 (19.74, 19.89)	**0.09 (0, 0.63)**	28.51 (0, 78.86)
10	517.19 (506.65, 527.74)	436.86 (433.39, 440.33)	450.06 (434.73, 465.39)	**410.11 (331.57, 488.66)**



(a) (b)

Figure 2. Cumulative operational cost and power unbalance for 10 days

VI. CONCLUSION

In this paper, the performance of policy-based DRL algorithms (DDPG, TD3, SAC, PPO) has been evaluated for the energy systems optimal scheduling problem. We provided a detailed sensitivity analysis for the penalty terms used to enforce the power balance constraint and minimize the operational cost in the reward function definition. Results showed that DRL algorithms are able to provide good quality solutions in real-time by exploiting the good generalization capabilities of deep learning models. In general, we observed that the PPO algorithm outperforms DDPG, SAC, and TD3 algorithms. The penalty coefficient used to balance the relationship between the operational cost and enforce the power balance constraint significantly impacted the performance of all DRL algorithms. Additionally, we observed that increasing the grid export/import power limit increases the flexibility of actions taken by algorithms (capability of providing feasible actions solutions). Nevertheless, all the tested DRL algorithms lack safety guarantees compared to model-based approaches, which could impede their practical implementation. In this regard, new approaches to enforce operational constraints in DRL algorithms are urgently needed.

ACKNOWLEDGMENT

This work made use of the Dutch national e-infrastructure with the support of the SURF Cooperative (grant no. EINF-2684) and was supported by the Chinese Scholarship Council (CSC) (grant no. 202106660002).

REFERENCES

[1] D. E. Olivares, C. A. Cañizares, and M. Kazerani, "A centralized optimal energy management system for microgrids," in *2011 IEEE Power and Energy Society General Meeting*. IEEE, 2011, pp. 1–6.

[2] M. F. Zia, E. Elbouchikhi, and M. Benbouzid, "Microgrids energy management systems: A critical review on methods, solutions, and prospects," *Applied energy*, vol. 222, pp. 1033–1055, 2018.

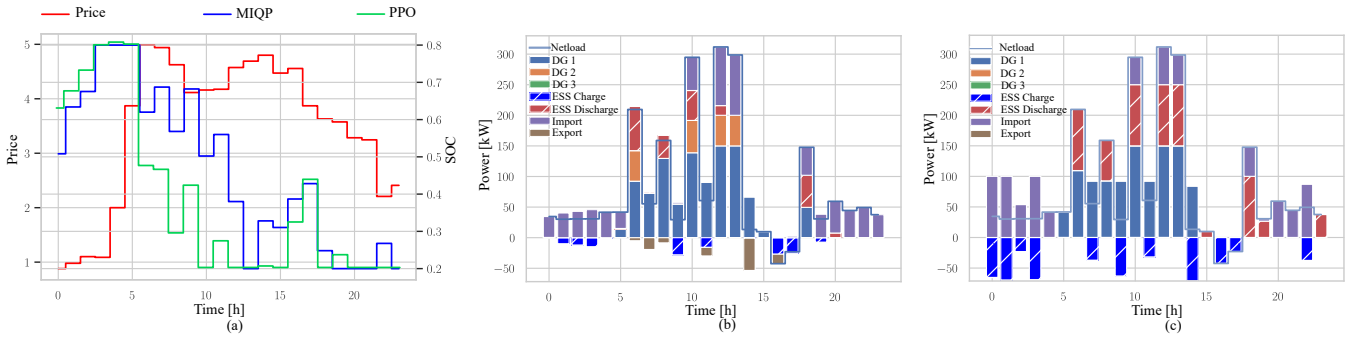


Figure 3. Operational schedule of all DG units and ESSs: (a) and (b) PPO algorithm, (b) and (c) MIQP formulation.

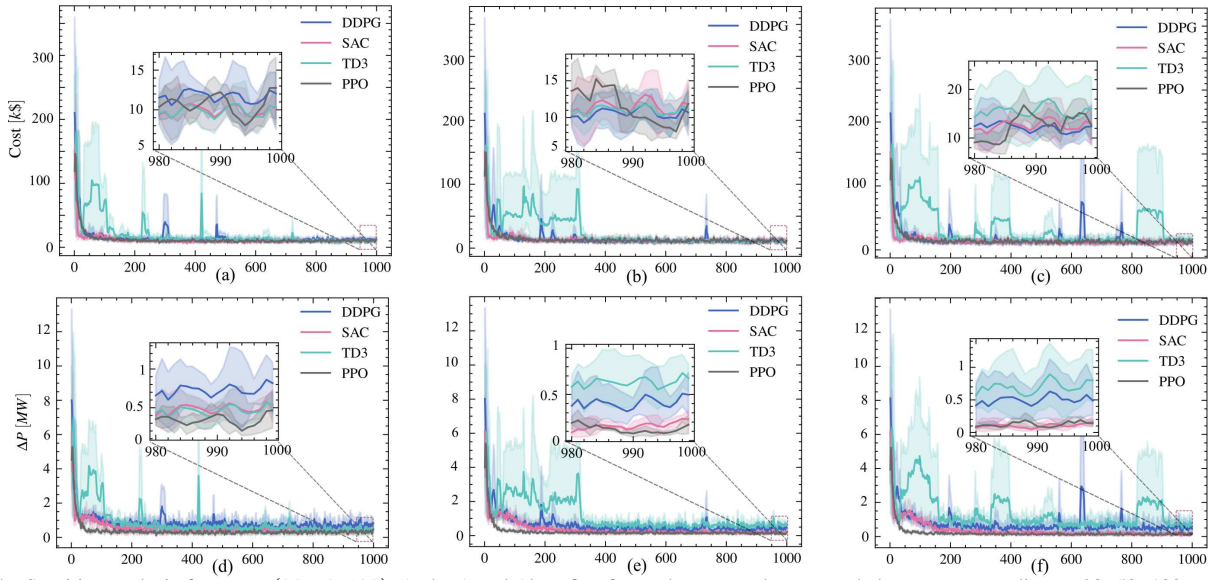


Figure 4. Sensitive analysis for $\sigma_2 = (20, 50, 100)$, (a, b, c) and (d, e, f) refer to the cost and power unbalance corresponding to 20, 50, 100, respectively.

[3] F. Hafiz, A. R. de Queiroz, P. Fajri, and I. Husain, "Energy management and optimal storage sizing for a shared community: A multi-stage stochastic programming approach," *Applied energy*, vol. 236, pp. 42–54, 2019.

[4] P. P. Vergara, M. Salazar, J. S. Giraldo, and P. Palensky, "Optimal dispatch of pv inverters in unbalanced distribution systems using reinforcement learning," *International Journal of Electrical Power & Energy Systems*, vol. 136, p. 107628, 2022.

[5] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for decision-making and control in power systems: Tutorial, review, and vision," *arXiv preprint arXiv:2102.01168*, 2021.

[6] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[7] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," *Sustainable Energy, Grids and Networks*, vol. 25, p. 100413, 2021.

[8] Y. Ji, J. Wang, J. Xu, X. Fang, and H. Zhang, "Real-time energy management of a microgrid using deep reinforcement learning," *Energies*, vol. 12, no. 12, p. 2291, 2019.

[9] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

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

[11] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[12] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[13] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[14] J. R. Vázquez-Canteli, S. Dey, G. Henze, and Z. Nagy, "Citylearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management," *arXiv preprint arXiv:2012.10504*, 2020.

[15] Y. Ji, J. Wang, J. Xu, and D. Li, "Data-driven online energy scheduling of a microgrid based on deep reinforcement learning," *Energies*, vol. 14, no. 8, p. 2120, 2021.

[16] S. Zhou, Z. Hu, W. Gu, M. Jiang, M. Chen, Q. Hong, and C. Booth, "Combined heat and power system intelligent economic dispatch: A deep reinforcement learning approach," *International journal of electrical power & energy systems*, vol. 120, p. 106016, 2020.

[17] H. Shengren, 2022, https://github.com/ShengrenHou/DRL-for-Energy-Systems-Optimal-Scheduling.

[18] W. E. Hart, C. D. Laird, J.-P. Watson, D. L. Woodruff, G. A. Hackebeil, B. L. Nicholson, J. D. Siirola *et al.*, *Pyomo-optimization modeling in python*. Springer, 2017, vol. 67.