Missouri University of Science and Technology
Scholars' Mine

Mathematics and Statistics Faculty Research & Creative Works

Mathematics and Statistics

01 Jun 2022

Joint Control of Manufacturing and Onsite Microgrid System Via Novel Neural-Network Integrated Reinforcement Learning Algorithms

Jiaojiao Yang

Zeyi Sun
Missouri University of Science and Technology, sunze@mst.edu

Wenqing Hu
Missouri University of Science and Technology, huwen@mst.edu

Louis Steinmeister

Follow this and additional works at: https://scholarsmine.mst.edu/math_stat_facwork

 Part of the Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation

J. Yang et al., "Joint Control of Manufacturing and Onsite Microgrid System Via Novel Neural-Network Integrated Reinforcement Learning Algorithms," *Applied Energy*, vol. 315, article no. 118982, Elsevier, Jun 2022.
The definitive version is available at https://doi.org/10.1016/j.apenergy.2022.118982

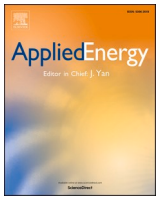
This Article - Journal is brought to you for free and open access by Scholars' Mine. It has been accepted for inclusion in Mathematics and Statistics Faculty Research & Creative Works by an authorized administrator of Scholars' Mine. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

Applied Energy 315 (2022) 118982

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Joint control of manufacturing and onsite microgrid system via novel neural-network integrated reinforcement learning algorithms

Jiaojiao Yang [a], Zeyi Sun [b,*], Wenqing Hu [c], Louis Steinmeister [c]

[a] School of Mathematics and Statistics, Anhui Normal University, Wuhu, Anhui 241002, China
[b] Mininglamp Technology, Shanghai 200030, China
[c] Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65401, USA

HIGHLIGHTS

- This paper proposes a joint energy control model for microgrid and manufacturing.
- Markov decision process is used to model the decision procedure.
- Reinforcement learning leveraging TD and DPG is proposed to solve the problem.

ARTICLE INFO

Keywords:
Microgrid
Manufacturing
Reinforcement Learning
Markov Decision Process
Temporal Difference Learning
Deterministic Policy Gradient

ABSTRACT

Microgrid is a promising technology of distributed energy supply system, which consists of storage devices, generation capacities including renewable sources, and controllable loads. It has been widely investigated and applied for residential and commercial end-use customers as well as critical facilities. In this paper, we propose a joint state-based dynamic control model on microgrids and manufacturing systems where optimal controls for both sides are implemented to coordinate the energy demand and supply so that the overall production cost can be minimized considering the constraint of production target. Markov Decision Process (MDP) is used to formulate the decision-making procedure. The main computing challenge to solve the formulated MDP lies in the co-existence of both discrete and continuous parts of the high-dimensional state/action space that are intertwined with constraints. A novel reinforcement learning algorithm that leverages both Temporal Difference (TD) and Deterministic Policy Gradient (DPG) algorithms is proposed to address the computation challenge. Experiments for a manufacturing system with an onsite microgrid system with renewable sources have been implemented to justify the effectiveness of the proposed method.

1. Introduction

A microgrid is a localized autonomous energy system that consists of distributed energy sources and loads [1], which can operate either separated from, or connected to, external utility power grids [2,3]. It is considered a reliable solution to satisfy the growing demand for electric power through strengthening the resilience and mitigating the disturbances of the grid [4,5].

Various studies on microgrids have been conducted for residential houses and critical facilities, such as medical centers, financial corporations, military bases, and jails (see details in Section 2.1). A great deal of literature focusing on the optimal control strategies of using microgrids has been reported (see details in Sections 2.2 and 2.3).

While for the manufacturing industry, the loss due to blackouts is not that direct or explicit when compared to the industries or sectors that are traditionally considered "critical". The incurred losses such as unemployment and supply chain failure cannot be sensed and evaluated immediately after blackouts. The overall influence in term of both economic and societal aspects seems to be underestimated. Thus, the research on the application of microgrid in manufacturing has been less reported compared to residential sector.

The benefits of combining microgrids with manufacturing systems can be discussed from two aspects. On one hand, the microgrids can offer one more option of energy supply in terms of cost effectiveness. The cost of energy delivered by utility company varies depending on different time periods. The use of microgrids can provide a more flexible

* Corresponding author.
E-mail address: sunzeyi@mininglamp.com (Z. Sun).

https://doi.org/10.1016/j.apenergy.2022.118982
Received 3 January 2022; Received in revised form 22 February 2022; Accepted 20 March 2022
Available online 4 April 2022
0306-2619/© 2022 Elsevier Ltd. All rights reserved.

solution of energy supply considering the cost variation. The users can select different energy sources as well as determine energy flow to optimize their performance with respect to cost effectiveness. In addition, the use of microgrids can enhance the security of manufacturing systems against the utility failures due to various reasons (e.g., natural disaster). On the other hand, microgrids can help manufacturing practitioners further improve their performance from the perspective of environmental sustainability. The adoption of renewable sources in microgrids can significantly improve the carbon footprints of the energy supplied for production.

Therefore, this study aims to enhance the use of microgrid in manufacturing sector through building a joint control model that can simultaneously adjust the energy supply and demand from both microgrid and manufacturing sides towards cost effectiveness and environment sustainability.

Recently, with the increasing concerns on climate change and environmental protection, the benefits of integrating energy-aware strategies in manufacturing have been recognized. Many sustainability-aware and carbon-constrained operating strategies, either on individual manufacturing processes or entire manufacturing system, have been investigated and used by many practitioners (see details in Section 2.4). Meanwhile, the use of microgrid to support manufacturing has also been launched, in particular, in the areas of benefits evaluation, system sizing, and microgrid-side control (see details in Sections 2.5 and 2.6), among which many investigations on using integrated energy systems for process industry have been reported (see details in Section 2.7).

However, the study of joint energy control for both microgrids and manufacturing systems simultaneously has not yet been fully launched. In this paper, a joint control model considering both microgrids and manufacturing systems is established using Markov Decision Process (MDP). A novel reinforcement learning algorithm is proposed for solving the MDP. The algorithm uses both Temporal Difference (TD) method to search for the discrete optimal control actions and Deterministic Policy Gradient (DPG) algorithms to search for the continuous optimal control actions, together with function approximations of the action-value function via a neural network. Experiments based on a manufacturing system with an onsite microgrid with renewable sources are implemented under real parameters to identify optimal control actions for both manufacturing system and microgrid towards cost optimality. It is empirically validated that the optimal policies found by the proposed reinforcement learning algorithm are more efficient in production and incur less cost when compared to randomly sampled policies and a routine operation policy. Last but not least, due to its exploration–exploitation nature, the proposed reinforcement learning method is also effective in the case when the MDP parameters are not known but have to be learned during the algorithm dynamics.

The proposed MDP model as well as the reinforcement learning algorithms can well address the major challenges with respect to modeling and problem solving. On one hand, the combined system including both manufacturing and microgrid has a complex interaction when controls are implemented from both sides. For example, the controls on manufacturing systems will influence the energy demand that needs to be met through controlling the operations of microgrids as well as the external utility connections to achieve an energy flow balance. The manufacturing system itself is a complex system where the interrelationships among different machines in the system need to be quantified. The manufacturing throughput should also be cared when energy control for the manufacturing system is implemented. All these factors need to be carefully considered.

On the other hand, it can be expected that the space of the states and the actions in the MDP model could be extremely large, which makes most existing strategies for solving MDP such as reinforcement learning algorithms less effective. An initial study by authors has shown that a traditional algorithm, e.g., vanilla Q-learning, integrated with a neural network can only work for a small sized model, while cannot sufficiently address the model with a large space size [6].

In summary, the major contributions of the proposed joint control model are summarized as follows:

- In most relevant literature, the control is only focused on the microgrid side, while neglecting the demand side (i.e., manufacturing system). In some research where manufacturing side is integrated into the control scheme, the manufacturing system is usually simplified and thus the complex internal dynamics cannot be fully represented (see details in Section 2.6). In this paper, the dynamic joint controls of manufacturing systems and microgrids, e.g., the complex dynamics in the demand side when energy control is implemented on manufacturing systems, are modeled.
- In existing literature, the renewable sources are usually modeled as the system state in MDP, that means, the renewable energy is dependent on the variations of wind and solar sources. It ignores the opportunities of controlling the energy supply from the renewable sources. In this paper, the availability of renewable sources is modeled as the state, while the ON/OFF control of renewable generation capacity is modeled as the control actions.
- There exist a few published works that use reinforcement learning algorithms such as deep Q-learning (DQN) to solve dynamic decision-making problems with respect to microgrid operations [7,8]. However, Q-learning algorithms is typically constrained by problem size, and thus cannot work very well if the problem size is too large. Our novel reinforcement learning algorithm integrates DPG with TD to treat the co-existence of discrete and continuous states and actions. We also address the constraints via proximal projection operators and policy gradient updates.

The remaining part of the paper is organized as follows. A brief literature review in several relevant areas is given in Section 2. Section 3 introduces the formulation of the dynamic decision-making model using MDP in detail. Section 4 proposes our reinforcement learning algorithm. Section 5 implements numerical case studies through extensive experiments and sensitivity analysis to validate the effectiveness of the proposed method. Section 6 concludes the paper and discusses future works. Supplementary materials in terms of mathematical derivations are summarized in Section 7.

2. Literature review

2.1. Microgrids for residential properties, community services, and critical facilities

The research of microgrid applications in residential properties, community services, and critical facilities has been widely reported in the literature. For example, Faruque discussed an economically profitable way to deploy a residential microgrid incorporating a new market entity that can act as both a consumer and a power supplier [9]. Kriett and Salani proposed a generic mixed integer linear programming model to minimize the operating cost of a residential microgrid [10]. Roggia et al designed a sustainable residential microgrid system including PHEV and energy storage devices [11]. Ahourai and Faruque analyzed the impact of a residential microgrid under various electric vehicle penetration levels [12]. Igualada et al. proposed an optimization model to manage a residential microgrid including a charging spot with a vehicle-to-grid system and renewable energy sources [13]. Hawkes and Leach investigated the cost-effective operating strategy for residential micro-combined heat and power (CHP) systems [14]. Tasdighi et al. investigated a micro-CHP based residential microgrid scheduling problem using smart meter data and temperature dependent thermal load modeling [15]. Kakigano et al. applied a direct-current microgrid to residential houses with a cogeneration system, such as a gas engine or fuel cell [16].

2.2. Microgrid control strategies using traditional methodologies

Considering the complex dynamics, in particular, the uncertainties of the renewable sources and the load, involved in microgrid operation, Model Predictive Control (MPC)-based approaches have been widely used to estimate the uncertainty and offer an optimizer to solve the best schedules of microgrid operations [17]. For example, an online optimal energy management model for energy storage system in a microgrid was developed using a mixed integer linear programming (MILP) over a rolling horizon period [18]. A similar method considering time-varying constraints of a microgrid was proposed [19]. MPC was applied to the operation of a hydrogen-based hybrid energy storage system in a microgrid [20]. A robust optimization model was proposed for microgrid operation using ensemble weather forecasts [21].

Another major research approach to address the challenges of uncertainties is stochastic optimization. For example, a stochastic energy scheduling model in microgrids with intermittent renewable energy resources was proposed [22]. A risk-averse stochastic programming method was proposed, which considered not only the expectation but the variation of the total cost [23]. A two-stage stochastic programming approach with MPC for microgrid energy management was proposed considering the uncertainty of load demand, renewable energy generation, and electricity prices [24]. From the perspective of control objectives, the existing literature focuses on optimizing cost [25,26], power system stability [27,28] such as voltage frequency control [29,30], and environmental objectives such as reducing carbon dioxide emissions [31,32].

2.3. Microgrid control strategies using Machine learning methodologies

Recently, some researchers have started to leverage the methodologies such as deep learning and reinforcement learning to address the microgrid control problem, i.e., reinforcement learning framework is used to solve the sequential decision-making problem typically formulated by MDP, and deep learning is used to approximate the values of some critical components in MDP such as Q-value and the state value. For example, Zeng et al. presented a novel dynamic energy management tool that incorporates efficient management of energy storage system into microgrid real-time dispatch while considering power flow constraints and uncertainties in load, renewable generation, and real-time electricity price [7]. Ji et al. proposed a deep reinforcement learning approach to identify the optimal control strategy of microgrids that can minimize the daily operating cost. In their study, a deep feed-forward neural network is designed to approximate the optimal action-value function, and the deep Q-network algorithm is used to train the neural network [8].

2.4. Sustainability-Aware and Carbon-Constrained manufacturing control strategy

Manufacturing activities dominate energy consumption and Greenhouse Gas (GHG) emissions in the industrial sector [33] that accounts for approximately one third of the total energy consumption in the U.S. [34]. Thus, many studies focusing on building energy consumption framework and exploring emission reductions for various industries have been reported. For example, a big data driven analytical framework was proposed to reduce the energy consumption and emission for energy-intensive industries [35,36]. Therblig power model was developed for calculating the energy supply of Computer Numerical Control (CNC) machine tools using machining process parameters [37].

In addition, many sustainability-aware and carbon-constrained operating and control strategies have been investigated for manufacturing. Optimal energy control towards a smart and sustainable manufacturing paradigm has been widely reported. For example, a CPS-enabled and knowledge-aided demand response strategy was proposed for sustainable manufacturing [38]. Later, the same team investigated a demand response strategy for manufacturing systems considering the implications of fast-charging battery powered material handling equipment [39]. Work-in-process parts in manufacturing systems were utilized for optimal production control to reduce energy cost without sacrificing production throughput [40]. A joint energy control and maintenance scheduling model was proposed to minimize overall operational costs considering time-dependent energy cost as well as equipment degradation [41].

2.5. Microgrid system sizing for manufacturing

The benefits of using microgrids in manufacturing have been gradually recognized by both academic colleagues and industrial practitioners. It is hardly possible to maintain manufacturing operations without electricity supply nowadays, even a very short power outage can lead to detrimental impacts on manufacturing companies [42,43]. Many studies focusing on the optimal design and component sizing of the microgrid for manufacturing plant has been reported [44–46]. For example, a Mixed Integer Non-Linear Programming optimization model was proposed for sizing the capacity of onsite generation system with renewable sources and battery energy storage system for the manufacturers considering the energy loads from both the manufacturing system and HVAC (Heating, Ventilation, and Air Conditioning) system in a typical manufacturing plant [44]. A time series model was proposed to describe and predict the variation of the energy load of manufacturing systems and the irradiation of solar energy such that a case study examining the cost for building and running an onsite microgrid system considering different microgrids capacities was implemented [46]. The onsite wind and solar power capacity that could maximize the cost saving of a manufacturing facility was estimated in an interruptible load demand response program [47].

2.6. Microgrid control for manufacturing

Some studies on using microgrid for manufacturing have been recently reported. For example, Harper et al. proposed a demand-side management method for manufacturing that uses a microgrid to reduce energy consumption and improve system reliability [48,49]. They employed a simple rule-based control strategy to adjust the operation modes of industrial pumps based on the empirical parameters of supply capability and electricity demand. Golari et al. presented a multi-period, production-inventory planning model in a multi-plant manufacturing system powered with onsite and grid renewable energy to determine the production quantity, the stock level, and the renewable energy supply in each period such that the aggregated production cost (including energy) is minimized [50]. It was formulated as a static optimization problem, rather than a dynamic state-based decision-making model.

The common limitations of the existing studies on microgrid control for manufacturing can be summarized as follows. 1) Manufacturing system is usually simplified and modeled as a single isolated machine, which ignores the complex dynamics and interactions of various equipment within the manufacturing system. 2) The manufacturing side is typically considered a fixed load, while neglecting its control flexibility. 3) The control strategies used for the manufacturing equipment are usually simple heuristic and/or empirical rule based utilizing the empirical parameters of supply capability and electricity demand.

2.7. Integrated energy systems in process industry

Process industries, e.g., food, pulp and paper, basic chemicals, refining, and iron/steel, are typically considered to be energy intensive. For example, the electricity demand of pulp & paper industry in 2020 was estimated to be 2.08EJ [51]. Many studies in terms of improving energy efficiency and reducing GHG emissions have been reported. For example, Pandey and Prakash systematically analyzed energy

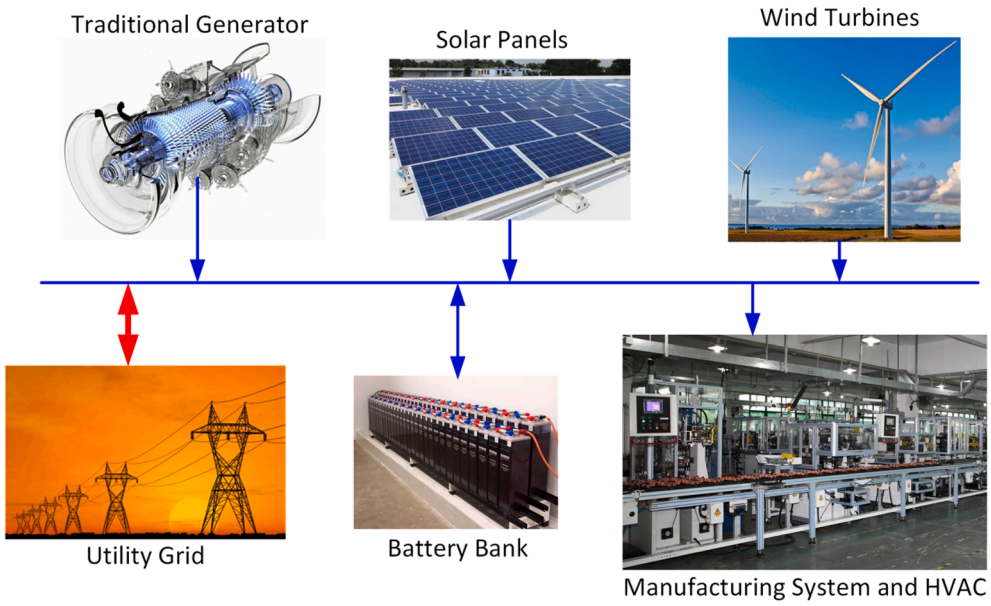


Fig. 1. A microgrid with various components.

Fig. 2. A typical manufacturing system with N machines and $N-1$ buffers (here $N = 5$).

conservation opportunities in pulp & paper industry [52]. Ashok proposed a peak load management model for steel plant [53]. Later, he and his colleague generalized the previous reseach to an optimization model for industrial load management [54]. Ma et al. presented an architecture of energy-cyber-physical system enabled management for energy-intensive process industries to enhance the implementation of integrated energy system for a cleaner production strategy [55]. Later, this team proposed a big data-driven analytical framework for energy-intensive process industries towards sustainability [56,57]. Zhang et al. explored the physical and chemical characteristics of fugitive emission sources in the iron and steel production process [58]. Sun et al. systematically analyzed the integrated optimization of material and energy flows in the iron and steel industry [59]. Zhang et al. proposed a carbon flow tracing and carbon accounting method for exploring CO_2 emissions of the iron and steel industry [60].

3. Markov decision process (MDP) for joint control of manufacturing and onsite microgrid systems

3.1. Formulation of joint energy control using MDP

3.1.1. Modeling microgrids and manufacturing systems

An MDP model is proposed to model the decision-making of the joint control of both onsite microgrid systems and manufacturing systems. MDP has been widely used in modeling complex and evolving state-based decision-making problems [61,62]. In this paper, the microgrid system used is a typical setup consisting of a gas turbine generator, a battery bank as well as solar PV modules and wind turbines as shown in Fig. 1. Note that, traditionally, microgrid meant a distributed energy generation system and renewable sources were not the default components. While, in recent years, with the increasing penetration and popularity of renewable sources in energy systems, the integration of wind and solar energy in microgrid has become the main stream in literature when studying microgrids, especially, considering the benefits with respect to environmental sustainability. Thus, renewable components are considered the default components in the microgrids modeled in this paper.

The manufacturing system modeled is a typical serial production line with N machines and $N-1$ buffers as shown in Fig. 2 (here $N = 5$, and work-in-progress parts are stored in buffers). Let $i = 1, \ldots, N$ be the indexes of the machines and $i = 1, \ldots, N-1$ be the indexes of the buffers.

Note that the type of the manufacturing system modeled in the paper is a flow shop setting. Many industries use such a layout for their production system, e.g., auto assembly system, aeroplane assembly system, etc. In practice, the system may be more complex, say, with some parallel configurations, while, a serial line is considered a fundamental layout from which many complex layouts can be studied through appropriate extensions.

The time horizon is discretized and divided into a set of discrete intervals, with the actual time duration for each interval to be Δt. The time variable t denotes the indexes of the decision epochs of such discrete intervals at which the control actions identified based on the optimal policy and the given states can be implemented. The state, policy, state transition, objective function, and constraints of the proposed MDP are introduced as follows.

3.1.2. System state

Let the system states form a state space S. The system state at decision epoch t is denoted by $\boldsymbol{S}_t$. It includes the states of manufacturing system ($\boldsymbol{S}_t^{mfg}$), microgrid system ($\boldsymbol{S}_t^{mic}$), and exogenous environmental features ($\boldsymbol{S}_t^{env}$), which can be formulated by $\boldsymbol{S}_t = \left(\boldsymbol{S}_t^{mfg}, \boldsymbol{S}_t^{mic}, \boldsymbol{S}_t^{env}\right)$. $\boldsymbol{S}_t^{mfg}$ can be denoted by $\boldsymbol{S}_t^{mfg} = \left(S_{1t}^M, \ldots, S_{Nt}^M, S_{1t}^B, \ldots, S_{(N-1)t}^B\right)$, where $S_{it}^M (i = 1, \ldots, N)$ denotes the state of machine i in the manufacturing system at decision epoch t; $S_{it}^B (i = 1, \ldots, N-1)$ denotes the state of the buffer i in the manufacturing system at decision epoch t. Machine states include operational, blockage, starvation, off, and breakdown. Blockage means that the machine itself is not failed while the completed part cannot be

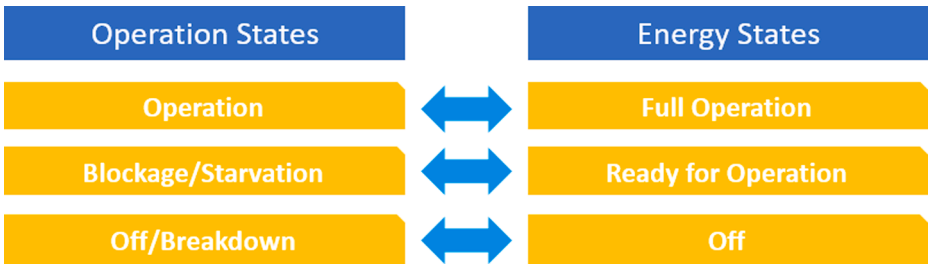


Fig. 3. Operation and energy states.

delivered to the downstream buffer due to the breakdown of specific downstream machines. Starvation means that the machine itself is not failed while there is no incoming part from the upstream buffer due to the breakdown of specific upstream machines. The set of machine states is thus $\{Opr, Blo, Sta, Off, Brk\}$, where *Opr*, *Blo*, *Sta*, *Off* and *Brk* denote the operational state, blockage state, starvation state, off state, and breakdown state, respectively. At each state, there is a corresponding power consumption state, which is illustrated in Fig. 3.

$\boldsymbol{S}_t^{mic}$ can be denoted by $\boldsymbol{S}_t^{mic} = (g_t^s, g_t^w, g_t^g, SOC_t)$, where g_t^s, g_t^w, and g_t^g denote the working status of solar PV, wind turbine, and generator, respectively, of the onsite microgrid generation system at decision epoch t (working $= 1$, not working $= 0$). A non-negative real number SOC_t, denotes the state of charge of the battery system at decision epoch t.

The state describing exogenous environmental features, $\boldsymbol{S}_t^{env}$, can be denoted by $\boldsymbol{S}_t^{env} = (I_t, v_t)$, where I_t denotes the solar irradiance at decision epoch t, and v_t denotes the wind speed at decision epoch t. The exogenous feature has an impact on the system dynamics and the cost function, but cannot be influenced by the control actions. The feature is time and weather dependent. The model formulation considers the availability of a deterministic forecast of the exogenous state information. The states (I_t, v_t) are taken from one year's data (assumed to be 360 days and 24 h/day, so in total 8640 h).

3.1.3. Control actions and policy

All admissible actions constitute an action space A**.** Let π be the policy that maps from different states $\boldsymbol{S}$ to the actions $\boldsymbol{A}$. The control actions adopted at decision epoch t can be denoted by $\boldsymbol{A}_t$. It includes the control actions for the manufacturing system $(\boldsymbol{A}_t^{mfg})$ and the microgrid system $(\boldsymbol{A}_t^{mic})$, which can be denoted by $\boldsymbol{A}_t = \left(\boldsymbol{A}_t^{mfg}, \boldsymbol{A}_t^{mic}\right)$. $\boldsymbol{A}_t^{mfg}$ can be denoted by $\boldsymbol{A}_t^{mfg} = \left(a_t^1, \ldots, a_t^N\right)$, where $a_t^i (i = 1, \ldots, N)$ is the control action for machine i at the decision epoch t. The actions include *K*-action, *W*-action, and *H*-action. *K*-action intends to keep original machine states, which can be applied to the machines in *Opr*, *Blo*, *Sta*, *Off*, and *Brk* states (note that machine repair is not considered a control action in this paper and repair is assumed to be a supposed-to-be reaction, so *K*-action used for breakdown machine can imply that repair will be implemented). *H*-action intends to turn off the machine, which can only be applied to the machine in *Opr*, *Blo*, and *Sta* states. *W*-action intends to turn on the machine that was previously turned off, which can only be applied to the machine in *Off* states. Note that for model simplicity, we assume that the energy consumption and time required for the transitions between different machine states can be ignored. The time and energy required for state transition depend on machine characteristics. For example, many CNC machines have a relatively short switch time as well as lower switch energy consumption when machine state is adjusted.

$\boldsymbol{A}_t^{mic}$ specifies the actions with respect to the adjustment of the working status of the components of the microgrid as well as the corresponding energy flow and allocation in the joint system, which can be denoted by $\boldsymbol{A}_t^{mic} = (a_t^s, a_t^w, a_t^g, s_t^m, s_t^b, s_t^{sb}, w_t^m, w_t^b, w_t^{sb}, g_t^m, g_t^b, g_t^{sb}, p_t^m, p_t^b, b_t^m)$. Here a_t^s, a_t^w, and a_t^g are the actions of adjusting the working status (i.e., 1 is connected, 0 is not connected to the load) of the solar, wind, and generators in microgrids; s_t^m, s_t^b, and s_t^{sb} denote the solar energy used for supporting manufacturing, charging battery, and sold back to grids, respectively. Similarly, the notations w and g with corresponding superscripts denote the allocation of the energy generated by wind turbine and generator, respectively. p_t^m and p_t^b denote the use of the energy purchased from the grid, i.e., for supporting manufacturing, and charging battery, respectively. Note that the energy purchased from the grid is not considered for sold back due to utility policies. Finally, b_t^m denotes the energy discharged by the battery for supporting manufacturing, and is given by a binary variable δ_t^{bm}, so that $b_t^m = b \cdot \delta_t^{bm} \cdot \Delta t$ for some discharging rate $b > 0$. Note that the energy discharged by the battery is not considered for sold back.

3.1.4. State transition

Let the function $P : \mathrm{S} \times \mathrm{A} \times \mathrm{S} \rightarrow [0, 1]$ be the transition probability function, so that $P(\boldsymbol{S}', \boldsymbol{A}, \boldsymbol{S}) \equiv Pr(\boldsymbol{S}' | \boldsymbol{S}, \boldsymbol{A})$ is the probability of transition to state $\boldsymbol{S}'$ given that the previous state was $\boldsymbol{S}$ and action $\boldsymbol{A}$ was taken. It is assumed that the state transition happens at the beginning of each interval when the decision is made.

For the manufacturing system, the buffer state at decision epoch $t+1$ can be obtained by (1) based on the states and the control actions adopted at decision epoch t of upstream and downstream machines:

$$S_{i(t+1)}^B = S_{it}^B + I\left(S_{it}^M, a_i^t\right) - I\left(S_{(i+1)t}^M, a_{i+1}^t\right), 0 \le S_{it}^B \le N_i, \tag{1}$$

Here N_i is the capacity of buffer i, and $I\left(S_{it}^M, a_i^t\right)$ is an indicator function that is defined by (2):

$$I\left(S_{it}^M, a_i^t\right) = \begin{cases} 1, when S_{it}^M = Opranda_i^t = K \\ 0\ , when S_{it}^M \neq Oprora_i^t = H \end{cases} \tag{2}$$

Referring to the literature focusing on the statistical methods for machine reliability [63], we assume L_i, which is the random lifetime of machine i, follows Weibull distribution with specific shape parameter and scale parameter. The probability that machine i goes into breakdown or non-breakdown state at the next decision epoch $t+1$, given it is not in breakdown state at the current decision epoch t can be described by (3) and (4) respectively:

$$Pr\left(S_{i(t+1)}^M = Brk | S_{it}^M \neq Brk, S_{it}^M \neq Off\right) = Pr(L_i < t + \Delta t) \tag{3}$$

$$Pr\left(S_{i(t+1)}^M \neq Brk | S_{it}^M \neq Brk, S_{it}^M \neq Off\right) = Pr(L_i \ge t + \Delta t) \tag{4}$$

Whether the machine is at *Off* state or not at next decision epoch can be determined in a deterministic way by (5):

$$S_{i(t+1)}^M = Off \text{ if } \left(S_{it}^M = Off \text{ and} a_i^t = K\right) \text{or} \left(S_{it}^M \neq Off \text{ and} a_i^t = H\right) \tag{5}$$

In addition, we also assume D_i, which is the random repair time of machine i, follows Exponential distribution [64]. The probability that machine i completes or does not complete the repair at the next decision epoch $t+1$, given it is in repair at the current decision epoch t can be described by (6) and (7) respectively.

$$Pr\left(S^M_{i(t+1)} \neq Brk | S^M_{it} = Brk\right) = Pr(D_i < t + \Delta t) \tag{6}$$

$$Pr\left(S^M_{i(t+1)} = Brk | S^M_{it} = Brk\right) = Pr(D_i \geq t + \Delta t) \tag{7}$$

From (3)-(7) we can decide whether the machine state $S^M_{i(t+1)}$ is *Brk* or *Off* or any one of the states *Sta*, *Blo*, *Opr*. To further determine the exact state if we fall into one of the states *Sta*, *Blo*, *Opr*, we use the following rules. For $2 \leq i \leq N$:

$$S^M_{i(t+1)} = Sta \text{if} S^M_{i(t+1)} \neq Brk/Off \text{and} S^B_{(i-1)(t+1)} = 0 \text{and} S^M_{(i-1)(t+1)} = Brk/Sta/Off \tag{8}$$

and for $1 \leq i \leq N-1$:

$$S^M_{i(t+1)} = Blo \text{if} S^M_{i(t+1)} \neq Brk/Off \text{and} S^B_{i(t+1)} = N_i \text{and} S^M_{(i+1)(t+1)} = Brk/Blo/Off \tag{9}$$

and for all $1 \leq i \leq N$:

$$S^M_{i(t+1)} = Opr \text{if} S^M_{i(t+1)} \neq Brk/Sta/Blo \tag{10}$$

Notice that (8) and (9) can be used to find all the *Sta* and *Blo* state machines by forward/backward scanning of all machines starting from the first/last machine, due to the trivial fact that $S^M_{1(t+1)} \neq Sta$ and $S^M_{N(t+1)} \neq Blo$. With these, (10) can further help to determine *Opr* state for machines.

Therefore, the system operation state transition between the current decision epoch and the next decision epoch can be calculated by using (1)-(10) when $\boldsymbol{A}^{mfg}_t$ is adopted based on a given $\boldsymbol{S}^{mfg}_t$.

For the microgrid system, the state transition of solar PV is determined by the action adopted. While, the state transition of wind turbine is determined by the action adopted and the variation of the wind speed. They can be formulated by (11) and (12), respectively:

$$g^s_{t+1} = \begin{cases} 1, if\, a^s_{t+1} = 1 \\ 0, if\, a^s_{t+1} = 0 \end{cases} \tag{11}$$

$$g^w_{t+1} = \begin{cases} 1, if\, a^w_{t+1} = 1 and v^{ci} \leq v_{t+1} \leq v^{co} \\ 0, if\, a^w_{t+1} = 0 or v_{t+1} > v^{co} or v_{t+1} < v^{ci} \end{cases} \tag{12}$$

where v^{ci} and v^{co} are the cut-in and cut-off wind speeds (m/s), respectively,

The state transition of generator is determined by the control actions adopted, which can be formulated by:

$$g^g_{t+1} = \begin{cases} 1, if\, a^g_{t+1} = 1 \\ 0, if\, a^g_{t+1} = 0 \end{cases} \tag{13}$$

The state transition of battery (i.e., *SOC*) is determined by the charging and discharging occurring between t and $t+1$ as well as the original *SOC*, which can be formulated by:

$$SOC_{t+1} = SOC_t + \left(s^b_t + w^b_t + g^b_t + p^b_t\right)\eta - b^m_t/\eta \tag{14}$$

where η is charging/discharging efficiency.

3.1.5. Objective function

The objective is to identify an optimal policy based on the given state that can minimize the incurred cost from time t to the end of decision horizon. At time t, the cost at state $\boldsymbol{S}_t$ when action $\boldsymbol{A}_t$ is taken, can be defined by $E(\boldsymbol{S}_t, \boldsymbol{A}_t)$. The total incurred cost from time 0 to the end of planning horizon, starting from state $\boldsymbol{S}$ and under policy $\pi : \mathrm{S} \rightarrow \mathrm{A}$, is given by:

$$C(\pi) = \mathbf{E}[\sum\nolimits_{t=0}^{\infty} \gamma^t E(\boldsymbol{S}_t, \pi(\boldsymbol{S}_t)) | \boldsymbol{S}_0 = \boldsymbol{S}] \tag{15}$$

Here $\gamma \in [0, 1)$ is the discount factor. The objective is to identify an optimal policy $\pi^* = \underset{\pi}{argmin} C(\pi)$ over all policies satisfying the constraints, that can guide the decision maker to find appropriate actions based on the given system state to minimize the total incurred cost $C(\pi)$ in (15).

For our model, the cost when action $\boldsymbol{A}$ is taken at state $\boldsymbol{S}$ is given by $E(\boldsymbol{S}, \boldsymbol{A})$, which is equal to energy consumption cost plus the microgrid operational cost, minus production throughput reward and the sold back reward. So, it can be calculated by:

$$E(\boldsymbol{S}, \boldsymbol{A}) = TF(\boldsymbol{S}, \boldsymbol{A}) + MC(\boldsymbol{S}, \boldsymbol{A}) - TP(\boldsymbol{S}, \boldsymbol{A}) - SB(\boldsymbol{S}, \boldsymbol{A}) \tag{16}$$

where $TF(\boldsymbol{S}, \boldsymbol{A})$ is the cost for the energy purchased from the grid, $MC(\boldsymbol{S}, \boldsymbol{A})$ is the operational cost for the onsite generation system, $TP(\boldsymbol{S}, \boldsymbol{A})$ is the reward of production throughput of the manufacturing system, and $SB(\boldsymbol{S}, \boldsymbol{A})$ is the sold back benefit.

Note that in this paper, since the throughput modeling and quantification for a typical manufacturing system is still a major research challenge in the field of production system engineering when considering non 100% reliable machines and finite capacity buffers [65,66], the concern of production throughput maintaining is represented as a monetary reward and integrated into the objective function. This strategy circumvents the challenges of modelling throughput as a major constraint in MDP.

In (16), $TF(\boldsymbol{S}, \boldsymbol{A})$ can be calculated by:

$$TF(\boldsymbol{S}, \boldsymbol{A}) = p_t \cdot r^c_t = \left(p^m_t + p^b_t\right) \cdot r^c_t \tag{17}$$

where p_t is the energy consumption purchased from the grid at decision epoch t, and r^c_t is the rate of energy consumption charge. p_t can be calculated by:

$$p_t = E^{mfg}_t - \left(s^m_t + w^m_t + g^m_t + b^m_t\right) \tag{18}$$

where E^{mfg}_t is the total energy consumed by the manufacturing system at decision epoch t which can be determined by:

$$E^{mfg}_t = \sum\nolimits_{i=1}^{N} PC_{it} \cdot \Delta t \tag{19}$$

where PC_{it} is the amount of power drawn by the machine i from t to $t+1$. PC_{it} can be calculated by:

$$PC_{it} = \begin{cases} 0, if S^M_{it} = Brk or S^M_{it} = Off \\ PC^{Opr}_i, if S^M_{it} = Opr \\ PC^{Idl}_i, if S^M_{it} = Sta or S^M_{it} = Blo \end{cases} \tag{20}$$

where PC^{Opr}_i and PC^{Idl}_i are the power level of machine i at the states of *Opr* and *Sta*/*Blo*, respectively.

$MC(\boldsymbol{S}, \boldsymbol{A})$ can be calculated by:

$$MC(\boldsymbol{S}, \boldsymbol{A}) = e^s_t \cdot r^s_{omc} + e^w_t \cdot r^w_{omc} + e^g_t \cdot r^g_{omc} + \frac{\left(b^m_t + s^b_t + w^b_t + g^b_t\right) \cdot \Delta t}{2e(SOC_{max} - SOC_{min})} \cdot r^b_{omc}, \tag{21}$$

where e^s_t, e^w_t, and e^g_t are the energy generated at decision epoch t from the onsite solar PV, wind turbine, and generator, respectively, which are calculated from the states and actions, that will be specified below. r^s_{omc}, r^w_{omc}, and r^g_{omc} are the unit operational and maintenance cost for generating electricity from solar PV, wind turbine, and generator respectively. r^b_{omc} is the operational and maintenance cost for battery storage system per unit charging/discharging cycle. e is the capacity of battery storage system. SOC_{max} and SOC_{min} are maximum and minimum allowed state of charge of battery storage system, respectively. Note that the fraction part of the 4th term on the right-hand side of (21) represents for the fraction of the charging/discharging cycle of the battery in one decision period [45]. Specifically, the numerator calculates the total energy flow (either charging or discharging) through the battery system in one decision period, while the denominator calculates the total energy flow of one charging/discharging cycle.

e^s_t can be calculated by:

$$e_t^s = \begin{cases} 0, if\, g_t^s = 0 \\ I_t \cdot a \cdot \delta \cdot \Delta t / 1000, if\, g_t^s = 1 \end{cases} \tag{22}$$

where I_t is the solar irradiance of a certain location (W/m^2) at decision epoch t, a is the area of the solar PV system, and δ is the efficiency of the solar PV.

e_t^w can be calculated by:

$$e_t^w = \begin{cases} 0, if\, g_t^w = 0 or v_t < v_{ci} or v_t > v_{co} \\ N_w \cdot RP_w \cdot \Delta t, if\, g_t^w = 1 and v^r \le v_t < v_{co} \\ N_w \cdot RP_w \cdot \dfrac{v_t - v_{ci}}{v^r - v_{ci}} \cdot \Delta t, if\, g_t^w = 1 and v_{ci} \le v_t < v^r \end{cases} \tag{23}$$

where v_t is the wind speed (m/s) at decision epoch t. v^r is the rated wind speeds (m/s). N_w is the number of wind turbine equipped within the onsite generation system and RP_w is the rated power of the wind turbine (kW). RP_w is determined by:

$$RP_w = \frac{1}{2} \cdot \rho \cdot \pi \cdot r^2 \cdot v_{avg}^3 \cdot \theta \cdot \eta_t \cdot \eta_g / 1000 \,, \tag{24}$$

where ρ is the density of air. v_{avg} is average wind speed. θ is the power coefficient. r is the radius of the wind turbine blade. η_t is its gearbox transmission efficiency. η_g is electrical generator efficiency.

e_t^g can be calculated by:

$$e_t^g = \begin{cases} 0, if\, g_t^g = 0 \\ n_g \cdot G_p \cdot \Delta t, if\, g_t^g = 1 \end{cases} \tag{25}$$

where n_g is the number of generators and G_p is the rated output power of the generator (kW).

$TP(\boldsymbol{S}, \boldsymbol{A})$ can be determined by:

$$TP(\boldsymbol{S}, \boldsymbol{A}) = pr_t \cdot r^p \tag{26}$$

where pr_t is the production count at decision epoch t and r^p is the unit reward for each unit of production. pr_t can be calculated by:

$$pr_t = \begin{cases} 1, if\, S_{Nt}^M = Opr and a_t^N = K \\ 0, if\, S_{Nt}^M \ne Opr or a_t^N = H \end{cases} \tag{27}$$

Note that, in such a serial production system as shown in Fig. 2, the system throughput is counted by the production count of the last machine in the line since one finished product is only counted when the process of the last machine is completed. Thus, the subscript in (27) is N.

Similarly, the sold back reward $SB(\boldsymbol{S}, \boldsymbol{A})$ can be calculated by:

$$SB(\boldsymbol{S}, \boldsymbol{A}) = s_t \cdot r^{sb} \,, \tag{28}$$

where $s_t = s_t^{sb} + w_t^{sb} + g_t^{sb}$ is the sold back energy to the grid at decision epoch t and r^{sb} is the unit reward from sold back energy.

3.2. Parameterization of the action space and constraints

Our model contains the following constraints for the action space A, that are described below.

Since we set that the battery state of charge level needs to be maintained within a given range, which can be formulated by:

$$SOC_{min} \le SOC_t + \left(s_t^b + w_t^b + g_t^b + p_t^b\right)\eta - \frac{b_t^m}{\eta} \le SOC_{max} \,. \tag{29}$$

Notice that (29) indicates that the actions to the microgrid on battery charging/discharging are controlled by the current SOC state.

Actions that can be applied to machines are restricted by the current machine states: K-action can be applied to the machine at Opr, Blo, Sta, Off and Brk states; H-action can only be applied to the machine at Opr, Blo and Sta states; W-action can only be applied to the machine at Off states, i.e.,

$$a_t^i = \begin{cases} K \,, if\, S_{it}^M = Opr, Blo, Sta, Off, Brk \\ H \,, if\, S_{it}^M = Opr, Blo, Sta \\ W \,, if\, S_{it}^M = Off \end{cases} \tag{30}$$

The energy flow balance for the energy generated by solar PV, wind turbine, and generator can be formulated by (31)-(33), respectively.

$$s_t^m + s_t^b + s_t^{sb} = e_t^s \,, \tag{31}$$

$$w_t^m + w_t^b + w_t^{sb} = e_t^w \tag{32}$$

$$g_t^m + g_t^b + g_t^{sb} = e_t^g \tag{33}$$

Notice that according to (22), (23) and (25), the energies e_t^s, e_t^w, e_t^g depend on the working states of the microgrid $\left(g_t^s, g_t^w, g_t^g\right)$, the solar irradiance I_t and the wind speed v_t. The constraints (31)-(33) are restricting the actions applied to microgrid based on the current microgrid state and the environmental features.

The battery cannot be charged and discharged simultaneously. The charge/discharge constraint is represented as follows:

$$\left(s_t^b + w_t^b + g_t^b + p_t^b\right) \cdot b_t^m = 0 \,. \tag{34}$$

Notice that since $b_t^m = \delta_t^{bm} \cdot b \cdot \Delta t$, we only seek for binary choices of $\delta_t^{bm} = 0/1$ when $s_t^b = w_t^b = g_t^b = p_t^b = 0$.

The energy sold back to the grid and the energy purchased from the grid cannot happen simultaneously, which can be represented by:

$$\left(s_t^{sb} + w_t^{sb} + g_t^{sb}\right)\left(p_t^m + p_t^b\right) = 0 \,. \tag{35}$$

If the constraint (35) is satisfied at $s_t^{sb} = w_t^{sb} = g_t^{sb} = 0$, so that we allow $p_t^m + p_t^b \ne 0$, then due to supply–demand balance principle, the energy purchased from the grid should be equal to the energy consumed from the grid, so we have.

$$p_t^m + p_t^b = p_t 1_{\{p_t > 0\}} \,, \tag{36}$$

where p_t is given by (18).

To simplify the model, we further assume that the energy purchased from the grid can only be used either for supporting manufacturing or charging battery, but not simultaneously, i.e.,

$$p_t^m \cdot p_t^b = 0 \,. \tag{37}$$

Due to (35) and (37), if $s_t^{sb} = w_t^{sb} = g_t^{sb} = 0$, we can introduce a binary variable $\delta_t^{pb} = 0/1$ (0 means purchased energy is not used for battery charging, 1 means purchased energy is used for battery charging) so that $p_t^m = \left(1 - \delta_t^{pb}\right) p_t 1_{\{p_t > 0\}}$ and $p_t^b = \delta_t^{pb} p_t 1_{\{p_t > 0\}}$; if else, i.e., any of s_t^{sb}, w_t^{sb} or g_t^{sb} is not equal to zero, we have $p_t^m = p_t^b = 0$.

To facilitate the design of policy-gradient related algorithms for training, we further parameterize the control actions $\left(s_t^m, s_t^b, s_t^{sb}, w_t^m, w_t^b, w_t^{sb}, g_t^m, g_t^b, g_t^{sb}\right)$ by introducing *proportionality parameters*.

$$\theta = \left(\lambda_s^m, \lambda_s^b, \lambda_w^m, \lambda_w^b, \lambda_g^m, \lambda_g^b\right) \tag{38}$$

such that the proportionality relation holds as.

$$\begin{cases} s_t^m = e_t^s \cdot \lambda_s^m \,, s_t^b = e_t^s \cdot \lambda_s^b \,, s_t^{sb} = e_t^s \cdot \left(1 - \lambda_s^m - \lambda_s^b\right) \\ w_t^m = e_t^w \cdot \lambda_w^m, w_t^b = e_t^w \cdot \lambda_w^b \,, w_t^{sb} = e_t^w \cdot \left(1 - \lambda_w^m - \lambda_w^b\right) \\ g_t^m = e_t^g \cdot \lambda_g^m \,, g_t^b = e_t^g \cdot \lambda_g^b \,, g_t^{sb} = e_t^g \cdot \left(1 - \lambda_g^m - \lambda_g^b\right) \end{cases} \tag{39}$$

These representations further simplify the constraints (31)-(33) into the following constraints.

$$\lambda_s^m \ge 0, \lambda_s^b \ge 0, 0 \le \lambda_s^m + \lambda_s^b \le 1 \tag{40}$$

$$\lambda_w^m \ge 0, \lambda_w^b \ge 0, 0 \le \lambda_w^m + \lambda_w^b \le 1 \tag{41}$$

$$\lambda_g^m \geq 0, \lambda_g^b \geq 0, 0 \leq \lambda_g^m + \lambda_g^b \leq 1 \tag{42}$$

To deal with the constraints (34), (35), and (37), we further introduce the binary (0/1) variables $\delta_t^b = 1_{\{s_t^b + w_t^b + g_t^b > 0\}}$, $\delta_t^{sb} = 1_{\{s_t^{sb} + w_t^{sb} + g_t^{sb} > 0\}}$, and $\delta_t^p = 1_{\{p_t^m + p_t^b > 0\}}$. Then constraints (34), (35), and (37) become a discrete constraint.

$$\left(\delta_t^b, \delta_t^p, \delta_t^{pb}, \delta_t^{sb}, \delta_t^{bm}\right) \in \{(1,1,1,0,0), (1,1,0,0,0), (1,0,0,1,0), (0,1,1,0,0), (0,1,0,0,1), (0,0,0,1,1)\}\,. \tag{43}$$

Notice that (43) summarizes all discrete constraints for the control parameters on the microgrid. The remaining continuous constraints for the microgrid are only (29) and (40)-(42). Based on (43), we can further write the constraints (29) and (40)-(42) into different constraints on θ (the parameter in (38)) and *SOC*. These constraints are formulated in (S1)-(S6) in Supplementary Material 7.3. All effective constraints for the admissible actions in this problem are (30), (43) and (S1)-(S6).

4. Novel Neural-Network integrated reinforcement learning algorithms for the MDP model

4.1. Abstract formulation of the MDP model

A state $\boldsymbol{S_t} \in \mathrm{S}$ in the space of the model consists of two parts $\boldsymbol{S_t} = \left(\boldsymbol{S_t^d}, \boldsymbol{S_t^c}\right)$: the discrete part.

$$\boldsymbol{S_t^d} = \left(S_{1t}^M, ..., S_{Nt}^M, S_{1t}^B, ..., S_{(N-1)t}^B, g_t^s, g_t^w, g_t^g, I_t, v_t\right) \tag{44}$$

which consists of the machine, buffer and microgrid states, as well as the coarse-grained solar irradiance and the wind speed (I_t, v_t). Here to reduce complexity, an approximate coarse-grained scheme is applied to each pair of the values (I_t, v_t), so that they will be approximated by integers closest to them on a grid with 20×20 states, thus taken values among 20×20 different states; the continuous part.

$$\boldsymbol{S_t^c} = (SOC_t) \tag{45}$$

which consists of the *SOC* state.

An action $\boldsymbol{A_t}$ in the action space A of the model also consists of three parts $\boldsymbol{A_t} = \left(\boldsymbol{A_t^d}, \boldsymbol{A_t^c}, \boldsymbol{A_t^r}\right)$: the discrete part.

$$\boldsymbol{A_t^d} = \left(a_t^1, ..., a_t^N, a_t^s, a_t^w, a_t^g, \delta_t^b, \delta_t^p, \delta_t^{pb}, \delta_t^{sb}, \delta_t^{bm}\right) \tag{46}$$

which consists of the actions on each machines and the connecting/disconnecting action of the solar PV, the wind turbine, and the generator, as well as the indicator variables in (43); the continuous part.

$$\boldsymbol{A_t^c} = \left(s_t^m, s_t^b, s_t^{sb}, w_t^m, w_t^b, w_t^{sb}, g_t^m, g_t^b, g_t^{sb}\right) \tag{47}$$

which consists of the solar/wind/generator energy used for supporting manufacturing, charging battery, and sold back to the grid. The continuous action $\boldsymbol{A_t^c}$ depends on $\boldsymbol{S_t}$ and the proportionality parameters in θ introduced in (38), i.e., $\boldsymbol{A_t^c} = \boldsymbol{A^c}(\theta_t, \boldsymbol{S_t})$; the remainder part $\boldsymbol{A_t^r} = \left(p_t^m, p_t^b, b_t^m\right)$, which consists of the use of the energy purchased from the grid for supporting manufacturing and charging battery, and the energy discharged by the battery for supporting manufacturing. These can be calculated directly from δ_t^{pb} and δ_t^{bm} as well as (36) and (18), which then can be calculated from $\boldsymbol{A_t^c}$.

At a specific state $\boldsymbol{S_t} \in \mathrm{S}$, the actions that can be taken are restricted by this particular state via the restriction.

$$\boldsymbol{A_t^d} \in D^d\left(\boldsymbol{S_t^d}\right) \tag{48}$$

And.

$$\theta \in D^c\left(\boldsymbol{S_t^d}, \boldsymbol{S_t^c}, \boldsymbol{A_t^d}\right) \tag{49}$$

where D^d is the set of admissible discrete actions $\boldsymbol{A_t^d}$ that can be taken at the current state, and D^c is the set of admissible parameters θ for continuous actions $\boldsymbol{A^c}$ that can be taken at the current state. According to the discussions in Section 3.2, we see that D^d is given by (30) and (43), and depends only on the discrete part of the current state (actually, only on the current states of the machines) and D^c is given by one of (S1)-(S6) in Supplementary Material 7.3, that depends on both the continuous and the discrete parts of the current state, as well as the discrete actions.

4.2. A brief overview of reinforcement learning: Temporal-Difference and policy gradient methods

For an infinite horizon, discounted MDP with state space S and action space A, Reinforcement Learning algorithms [67] design an "agent" interacting with an environment over a number of time steps. At each time step t when the agent is at a state $\boldsymbol{S_t}$, it selects an action $\boldsymbol{A_t}$ using the policy $\pi(\boldsymbol{S_t}) \in \mathrm{A}$, and then moves to the next state $\boldsymbol{S_{t+1}}$ according to the transition probability $P(\boldsymbol{S_{t+1}}, \boldsymbol{A_t}, \boldsymbol{S_t}) = Pr(\boldsymbol{S_{t+1}}|\boldsymbol{S_t}, \boldsymbol{A_t})$. During this step the agent incurs a cost $E(\boldsymbol{S_t}, \boldsymbol{A_t})$. The goal of the agent is to find a policy π^* that minimizes the expected total discounted cost.

$$\pi^* = \underset{\pi}{argmin} C(\pi) \tag{50}$$

where $C(\pi) \equiv \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t E(\boldsymbol{S_t}, \boldsymbol{A_t}) | \boldsymbol{S_0} = \boldsymbol{S}\right]$, and $\gamma \in (0, 1]$ is the discount factor. To solve (50), we introduce the *action-value function* $Q^\pi(\boldsymbol{S}, \boldsymbol{A})$, which is the expected total discounted cost under policy π starting from state $\boldsymbol{S}$ and taking initial action $\boldsymbol{A}$. For the optimal policy π^* in (50) we have $Q^*(\boldsymbol{S}, \boldsymbol{A}) \equiv Q^{\pi^*}(\boldsymbol{S}, \boldsymbol{A}) = min_\pi Q^\pi(\boldsymbol{S}, \boldsymbol{A})$. The key idea in action-value methods is to identify π^* through a sequence of approximators $Q(\boldsymbol{S}, \boldsymbol{A})$ built via an iterative scheme that approaches $Q^*(\boldsymbol{S}, \boldsymbol{A})$. In practice we use a function approximator, such as a neural network $Q(\boldsymbol{S}, \boldsymbol{A}; \omega)$ (sometimes referred as the "critic", see [67–70]) to replace the actual action-value function $Q^\pi(\boldsymbol{S}, \boldsymbol{A})$, where ω is the parameter of the approximating function. Given current ω_t, we construct the *temporal difference*.

$$\delta_t = E(\boldsymbol{S_t}, \boldsymbol{A_t}) + \gamma Q(\boldsymbol{S_{t+1}}, \boldsymbol{A_{t+1}}; \omega_t) - Q(\boldsymbol{S_t}, \boldsymbol{A_t}; \omega_t) \tag{51}$$

where $\boldsymbol{S_{t+1}}$ is the next state of MDP from state $\boldsymbol{S_t}$ under action $\boldsymbol{A_t}$ and $\boldsymbol{A_{t+1}}$ is selected as the optimal policy (or near-optimal policy), such as ε-greedy policy [67] at state $\boldsymbol{S_{t+1}}$ when the approximator is given by the parameter ω_t. This class of action-value methods are called (on-policy) *Temporal Differences* (TD or SARSA, see [67]). Actually, there are two kinds of reinforcement learning methods: on-policy and off-policy. On-policy methods attempt to evaluate or improve the policy that is used to make decisions, whereas off-policy methods evaluate or improve a policy different from that used to generate the data (see [67, Section 5.4, p. 82]). In our temporal-difference method, we are using the policy obtained from data in the current iteration step to generate new data and we repeat this iteration. So, our TD method is on-policy. Once the temporal differences (51) are constructed, we can update the critic parameter ω_t according to a standard gradient descent.

$$\omega_{t+1} = \omega_t - \eta_\omega \delta_t \nabla_\omega Q(\boldsymbol{S_t}, \boldsymbol{A_t}; \omega_t)\,, \tag{52}$$

where $\eta_\omega > 0$ is the learning rate. Another way to solve (50) works in the scenario when the policy can be parameterized by a policy

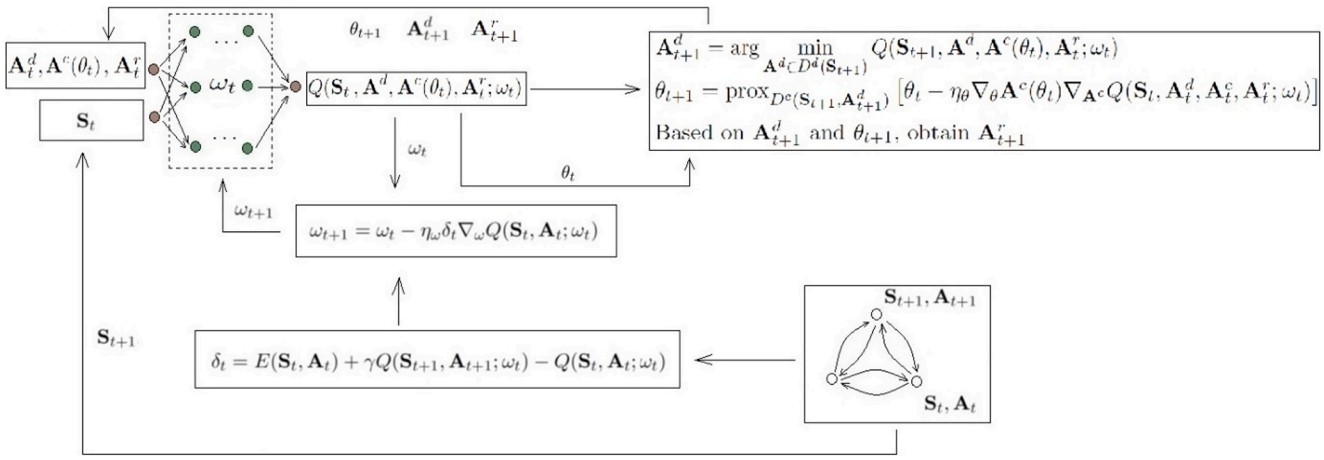


Fig. 4. The flow chart of Algorithm 1.

parameter θ, so that $\pi = \pi_\theta$, $C(\pi) = C(\pi_\theta) = C(\theta)$. In that case, we can simply use a gradient descent to find the optimal policy parameter θ^* such that $\pi^* = \pi_{\theta^*}$:

$$\theta_{t+1} = \theta_t - \eta_\theta \nabla_\theta C(\theta) \tag{53}$$

where $\eta_\theta > 0$ is the learning rate. The gradient $\nabla_\theta C(\theta)$ can be estimated by the well-celebrated *policy gradient theorem* when the policy π_θ is random (see [67]). However, if we are restricted only to deterministic policies, so that we assume that $\pi_\theta(\boldsymbol{S}) = \boldsymbol{A}(\theta)$ is a deterministic policy, and thus $C(\theta) = C(\boldsymbol{A}(\theta))$, then the Deterministic Policy Gradient (DPG, see [68]) can also be calculated as:

$$\nabla_\theta C(\theta) = \mathbf{E}_{S\sim\rho^\pi|_{\pi=\boldsymbol{A}(\theta)}} \left[\nabla_\theta \boldsymbol{A}(\theta) \nabla_{\boldsymbol{A}} Q^\pi(\boldsymbol{S}, \boldsymbol{A})|_{\pi=\boldsymbol{A}(\theta), \boldsymbol{A}=\boldsymbol{A}(\theta)} \right] \tag{54}$$

where $\rho^\pi|_{\pi=\boldsymbol{A}(\theta)} = \mathbf{E}_{S_0} \sum_{t=1}^\infty \gamma^t Pr(\boldsymbol{S}_t = \boldsymbol{S}|S_0, \pi)$ is the discounted state distribution when $\pi = \boldsymbol{A}(\theta)$.

4.3. Neural-Network integrated reinforcement learning for solving the optimal policy of the model in its abstract formulation

Based on the abstract formulation in Section 4.1, a policy π in the objective function (15) can be written as $\pi(\boldsymbol{S}_t) = (\boldsymbol{A}_t^{\boldsymbol{d}}(\boldsymbol{S}_t), \boldsymbol{A}^{\boldsymbol{c}}(\theta_t, \boldsymbol{S}_t), \boldsymbol{A}_t^r(\boldsymbol{S}_t))$. The purpose is then to find optimal proportionality parameter $\boldsymbol{\theta}^*$ and for each state $\boldsymbol{S} \in$ S, to identify the optimal control actions $(\boldsymbol{A}^{*,\boldsymbol{d}}(\boldsymbol{S}), \boldsymbol{A}^{*,\boldsymbol{c}}(\theta^*, \boldsymbol{S}), \boldsymbol{A}^{*,r}(\boldsymbol{S}))$, so that the objective function (15) can be minimized. Notice that, if θ is the only parameter for the control actions, then DPG can be used to search for an optimal θ. On the other hand, if we only have to search for the optimal discrete actions $\boldsymbol{A}^{*,\boldsymbol{d}}(\boldsymbol{S})$, then we can employ TD. Since we are simultaneously targeting at these two tasks, we alternatively perform these two optimization steps. We take into account the constraints $D(\boldsymbol{S})$, so that at the TD step, we search over the constraint $D^d(\boldsymbol{S})$ on the discrete action space $\boldsymbol{A}^d$, while at the policy gradient step, we project the gradient updates to the continuous constraint $D^c(\boldsymbol{S})$ on $\boldsymbol{A}^c$ using proximal projection operators (see [72]). We employ on-policy methods rather than off-policy to deal with constraints that are variable with respect to change of states. This enables more exploration over the variable constraints.

The proposed method is model-free (see [67]), as compared with model-based methods (see [73]). This means that we do not have to know in advance any of the MDP system parameters. Rather, the exploration mechanism in the reinforcement learning algorithm will help us to estimate all the system parameters in an implicit way and make use of them. In the case study in Section 5, the model parameters are given only in order to simulate the system, but the learning algorithm does not require to know these parameters.

Below we write the proposed algorithm into pseudo-code at Algorithm 1.

Algorithm 1. *Reinforcement Learning via integrating TD control and DPG with Proximal Projection.*

1: **Input**: Input state space S, action space A, constraints D^d,D^c; Given the neural network architecture $Q(\boldsymbol{S}, \boldsymbol{A}^d, \boldsymbol{A}^c(\theta), \boldsymbol{A}^r; \omega)$; Discount factor $0 < \gamma < 1$; Learning rates $\eta_\theta, \eta_\omega > 0$;

2: **Initialization**: Initialize the weight vector $\omega_0 \sim$ agivenpriordistribution; initial action $\boldsymbol{A}_0^d, \boldsymbol{A}_0^r$ and action parameter θ_0, $\boldsymbol{A}_0^c = \boldsymbol{A}^c(\theta_0)$; Initial state $\mathrm{S}_0 = (\mathrm{S}_0^d, \mathrm{S}_0^c)$;

3: **for** $t = 1, 2, \cdots, T$ **do**.

4: Run one step of the MDP from state S_t under action $\boldsymbol{A}_t = (\boldsymbol{A}_t^d, \boldsymbol{A}_t^c, \boldsymbol{A}_t^r)$, obtain a new state S_{t+1};

5: Calculate the total cost $E(\boldsymbol{S}_t, \boldsymbol{A}_t)$;

6: Identify.

$$\boldsymbol{A}_{t+1}^d = \arg \min_{\boldsymbol{A}^d \in D^d(\mathrm{S}_{t+1})} Q(\boldsymbol{S}_{t+1}, \boldsymbol{A}^d, \boldsymbol{A}^c(\theta_t), \boldsymbol{A}_t^r; \omega_t) ;$$

7: Based on $\boldsymbol{A}_{t+1}^d$, update the policy parameter θ according to DPG (53) and (54):

$$\theta_{t+1} = \text{prox}_{D^c(\mathrm{S}_{t+1}, \boldsymbol{A}_{t+1}^d)} [\theta_t - \eta_\theta \nabla_\theta \boldsymbol{A}^c(\theta_t) \nabla_{\boldsymbol{A}^c} Q(\boldsymbol{S}_t, \boldsymbol{A}_t^d, \boldsymbol{A}_t^c, \boldsymbol{A}_t^r; \omega_t)] ;$$

8: Based on $\boldsymbol{A}_{t+1}^d$ and θ_{t+1}, obtain $\boldsymbol{A}_{t+1}^r$, so that.

$$\boldsymbol{A}_{t+1} = (\boldsymbol{A}_{t+1}^d, \boldsymbol{A}_{t+1}^c = \boldsymbol{A}^c(\theta_{t+1}), \boldsymbol{A}_{t+1}^r) ;$$

9: Calculate TD (51):

$$\delta_t = E(\boldsymbol{S}_t, \boldsymbol{A}_t) + \gamma Q(\boldsymbol{S}_{t+1}, \boldsymbol{A}_{t+1}; \omega_t) - Q(\boldsymbol{S}_t, \boldsymbol{A}_t; \omega_t) ;$$

10: Update the weight vector ω of the critic based on (52):

$$\omega_{t+1} = \omega_t - \eta_\omega \delta_t \nabla_\omega Q(\boldsymbol{S}_t, \boldsymbol{A}_t; \omega_t) ;$$

11: **end for**.

12: **Output**: With the given optimal ω^* and θ^* found, for each state $\boldsymbol{S}$ given, output the approximate optimal policy $(\boldsymbol{A}^{*,d}, \boldsymbol{A}^c(\theta^*), \boldsymbol{A}^{*,r})$, where.

$$(\boldsymbol{A}^{*,d}, \boldsymbol{A}^{*,r}) = \arg \min_{\boldsymbol{A}^d, \boldsymbol{A}^r \text{admissible}} Q(\boldsymbol{S}, \boldsymbol{A}^d, \boldsymbol{A}^c(\theta^*), \boldsymbol{A}^r; \omega^*) .$$

The flow chart of Algorithm 1 is shown in Fig. 4 below.

4.4. Practical implementation details of Algorithm 1 at original MDP model

When implementing Algorithm 1 at the original MDP model for the microgrid-manufacturing system, practical issues need to be addressed

Table 1
Machine Parameters.

	Mean time between failures Scale/Shape Parameter	Mean time to repair (min)	Rated power of operation (kW)	Power at idle state (kW)
M_1	111.39 min/1.5766	4.95	115.5	105
M_2	51.1 min/1.6532	11.7	115.5	105
M_3	110.9 min/1.7174	15.97	115.5	105
M_4	239.1 min/1.421	27.28	170.5	155
M_5	122.1 min/1.591	18.37	132	120

Table 2
Buffer Parameters.

	B_1	B_2	B_3	B_4
Capacity	1000	1000	1000	1000
Initial	100	100	100	100

arise. The set of admissible parameters $\theta \in D^c(\boldsymbol{S},\boldsymbol{A})$ for continuous actions $\boldsymbol{A^c}$, that are determined by (S1)-(S6) in Supplementary Material 7.3, have an intersection structure. Indeed from (S1)-(S6) one can view $D^c(\boldsymbol{S},\boldsymbol{A})$ as an intersection $D^c(\boldsymbol{S},\boldsymbol{A}) = D^c \cap D^{SOC}\left(\theta, SOC_t, \left(\delta_t^b, \delta_t^p, \delta_t^{pb}, \delta_t^{sb}, \delta_t^{bm}\right)\right)$. Here we set the fixed simplex.

Table 3
Wind Turbine Parameters.

Parameters	Value	Parameters	Value
$v^{ci}(m/s)$-Cut in speed	3	η_t-Gearbox efficiency	0.9
$v^{co}(m/s)$-Cut off speed	11	η_g-Generator efficiency	0.9
$v^r(m/s)$-Rate speed	7	θ-Power coefficient	0.593
$\rho(kg/m^3)$-Air density	1.225	r^w_{omc}(\$/kWh)	0.08
r(m)-Blade radius	25	N_w(unit)	1

Table 4
Battery Storage Parameters.

Parameters	Value	Parameters	Value
e(kWh)-Capacity	350	b-Charging rate (kW)	2
SOC_{max}(%)	95	SOC_{min}(%)	5
r^b_{omc}(\$/kWh)	0.9	η-Efficiency	0.99

- Once θ is chosen, we determine whether or not we should choose $p^m_{t+1} + p^b_{t+1} \neq 0$, and whether $p^b_{t+1} \neq 0$ according to (35), (36), and (37); we also determine whether or not we choose $b^m_{t+1} \neq 0$ according to (34). The precise scheme of choosing $(p^m_{t+1}, p^b_{t+1}, b^m_{t+1})$ can be found in Supplementary Material 7.2.
- With all these ready, Step 8 in Algorithm 1 will be modified accordingly.

$$D^c = \{\left(\lambda_s^m, \lambda_s^b, \lambda_w^m, \lambda_w^b, \lambda_g^m, \lambda_g^b\right) : \lambda_s^m, \lambda_s^b, \lambda_w^m, \lambda_w^b, \lambda_g^m, \lambda_g^b \geq 0, 0 \leq \lambda_g^m + \lambda_g^b \leq 1, 0 \leq \lambda_s^m + \lambda_s^b \leq 1\ , 0 \leq \lambda_w^m + \lambda_w^b \leq 1\} \tag{55}$$

The complicacy in $\mathbf{D^c}(\boldsymbol{S},\boldsymbol{A})$ lies in the other part of the intersection $\mathbf{D^{SOC}}\left(\theta, SOC_t, \left(\delta_t^b, \delta_t^p, \delta_t^{pb}, \delta_t^{sb}, \delta_t^{bm}\right)\right)$, that may vary according to the choice of $\left(\delta_t^b, \delta_t^p, \delta_t^{pb}, \delta_t^{sb}, \delta_t^{bm}\right)$ and depend on the *SOC* state, which makes the proximal projection to $D^c(\boldsymbol{S},\boldsymbol{A})$ in Step 7 of Algorithm 1 nearly impossible to compute in practice. To fix this issue, we suggest to *relax* the constraint $D^c(\boldsymbol{S},\boldsymbol{A})$ by only considering its fixed simplex part D^c. Of course, if we only project to D^c at every proximal projection step in our Step 7 of Algorithm 1, we may miss the *SOC* constraints in (S1)-(S6). But we can then fix this issue by using the θ found on the relaxed constraint set D^c and determine the binary variables $\left(\delta_t^b, \delta_t^p, \delta_t^{pb}, \delta_t^{sb}, \delta_t^{bm}\right)$, as well as update the *SOC*. If the *SOC* values we obtain violate the additional constraints in $D^{SOC}\left(\theta, SOC_t, \left(\delta_t^b, \delta_t^p, \delta_t^{pb}, \delta_t^{sb}, \delta_t^{bm}\right)\right)$, we will just set them to be the boundary values SOC_{max} or SOC_{min}, so that they will not violate the *SOC* constraints. In this way, we can find an approximate solution to the optimal one, with computationally achievable simulations.

Under this simplification, the solution algorithm for the original model becomes practical to implement. We can then change Algorithm 1 accordingly, so that.

- In Step 6 of Algorithm 1, we select one action A^d_{t+1} that only takes into account the constraint (30) and minimizes the Q-value. The admissible actions $\boldsymbol{A^d}$ are stored in a tree constructed based on (30) and we search over the tree to identify optimal action A^d_{t+1} based on Step 6. The action A^d_{t+1} does not include the indicator variables $\left(\delta^b_{t+1}, \delta^p_{t+1}, \delta^{pb}_{t+1}, \delta^{sb}_{t+1}, \delta^{bm}_{t+1}\right)$ in (43).
- In Step 7 of Algorithm 1, the constraint $D^c\left(\boldsymbol{S}_{t+1}, \boldsymbol{A}^d_{t+1}\right)$ is replaced by a fixed constraint D^c in (55). Notice that in this case, the projection onto D^c can be calculated directly, see Supplementary Material 7.1.
- At each loop, the optimal actions are implemented for the MDP system to jump to the next state and the incurred cost, throughput and energy demand are calculated.

5. Case study

5.1. Experimental parameters

5.1.1. Joint Microgrid-Manufacturing system parameters

In this section, numerical case studies are implemented to illustrate the benefits of the proposed modeling and solution strategies. The open-source code for the proposed model in this paper is released at the GitHub repository [74].

We have carried out experiments of the microgrid-manufacturing

Table 5
Solar Panel and Generator Parameters.

Parameters	Value	Parameters	Value
$a(m^2)$-Panel area	1400	n_g(unit)-Number of generator	1
δ-Efficiency	0.2	G_p(kW)-Generator capacity	650
r^s_{omc}(\$/kWh)	0.17	r^g_{omc}(\$/kWh)-Operating cost	0.45

Table 6
Comparison Among Three Models.

	Energy Cost (\$)	Production Throughput (unit)	Training CPU Time (s)
Routine Strategy	3,642	73	29
Proposed Model	876	73	457
Random Policy	1,656	24	N/A

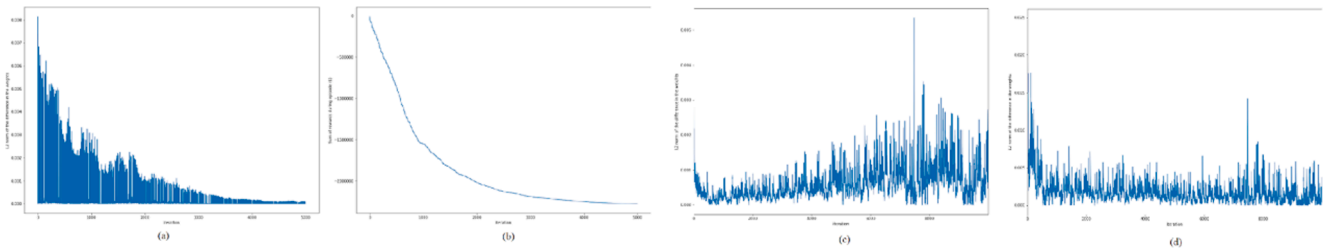


Fig. 5. Left to Right: (a) Evolution of weight differences $\|\omega_{t+1}-\omega_t\|_2^2$; (b) Evolution of cumulative rewards; (c), (d) Compare with vanilla Q-learning.

system using a real-case parameter set. The manufacturing system includes five machines and four buffers as shown in Fig. 2. We summarize all the system parameters in Tables 1-5. The parameters of the manufacturing system are taken according to [78], where machine and buffer parameters of the manufacturing system are shown in Table 1 and Table 2, respectively. Note that the mean time between failures of each machine is modeled as random variables following Weibull distribution with respective scale and shape parameters. The mean time to repair of each machine is modeled as exponentially distributed random variables. Unit production reward r^p is set to be $\$10^4$ per unit.

The parameters of the microgrid used in the experiment are sized based on the manufacturing load according to the methods proposed in [45]. The parameters related to wind turbine, battery storage system, and solar panel and generator are illustrated in Table 4, Table 5, and Table 6, respectively. The data of solar irradiance and wind speed are collected from Solar Energy Local [76] and State Climatologist of Illinois [77], respectively.

In order to prevent computational overflow, in the actual numerical experiment we scaled all the parameters by choosing different units of measurement, with distance measured by km (10^3 m), time measured by hour (60 min = 3600 s), speed measured by km/h (3.6 m/s), energy measured by MegaWatt (10^6 Watt), money cost measured by 10^4 USD, area measured by km^2 (10^6 m^2), mass measured by 10^6 kg. Time period is measured in hours.

5.1.2. Reinforcement algorithm parameters

The neural network $Q(\boldsymbol{S},\boldsymbol{A};\omega)$ is taken to be fully connected and it contains two hidden layers with 100 neurons for each layer, with Sigmoid and ReLU activations for layers 1 and 2. The output is then scaled back to usual units of measurement, like $ for cost and kW for energy demand.

We take the discount factor $\gamma = 0.999$. The experiment includes reinforcement learning training for 5×10^3 iterations, with each iteration counts for time period with equal duration of one hour per period, and the learning rates are tuned to be $\eta_\theta = 0.003$ and $\eta_\omega = 0.0003$, with η_ω discounted by a factor 0.999 at each iteration. Fig. 5-(a) plots the L^2-norm of the difference at two consecutive iterations during training in the neural network weights ($\|\omega_{t+1}-\omega_t\|_2^2$), which clearly indicates the convergence of the training process. Fig. 5-(b) plots the cumulative reward function (total incurred cost $C(\boldsymbol{S},\pi)$ in (15) truncated at the current iteration step).

In order to validate the effective convergence of our reinforcement learning algorithm, comparison has been carried out for a pure Q-learning algorithm for the same microgrid-manufacturing system with a smaller size (2 machines and 1 buffer) [6]. After discretization of the continuous states and actions, the state space has a size of 3.8×10^3 and the action space has a size of 2.6×10^4. We use a smaller fully-connected neural network $Q(\boldsymbol{S},\boldsymbol{A};\omega)$ with two hidden layers and Sigmoid activation, where each hidden layer has 32 neurons. Again, we calculated the square norms of the differences of the neural-network weight vectors for each two consecutive algorithm iterations (i.e., $\|\omega_{t+1}-\omega_t\|_2^2$). The discount factor $\gamma = 0.1$. The neural network is trained using Adam [75] with different learning rates: Fig. 5-(c) is for learning rate 0.001 and Fig. 5-(d) is for learning rate 0.0001. It is seen that in these two cases, even after 10^4 iterations, the vanilla Q-learning algorithm cannot converge due to the immense size of the discretized state-action spaces (a manifestation of the "curse of dimensionality"), indicating the effectiveness of our method that combines deterministic policy gradient [71] with discrete Monte-Carlo type searches. Based on the optimal parameters (ω^*,θ^*) found for the neural-network $Q(\boldsymbol{S},\boldsymbol{A};\omega^*)$ and the continuous action $\boldsymbol{A}^c(\theta^*)$ (see Algorithm 1), we tested the corresponding

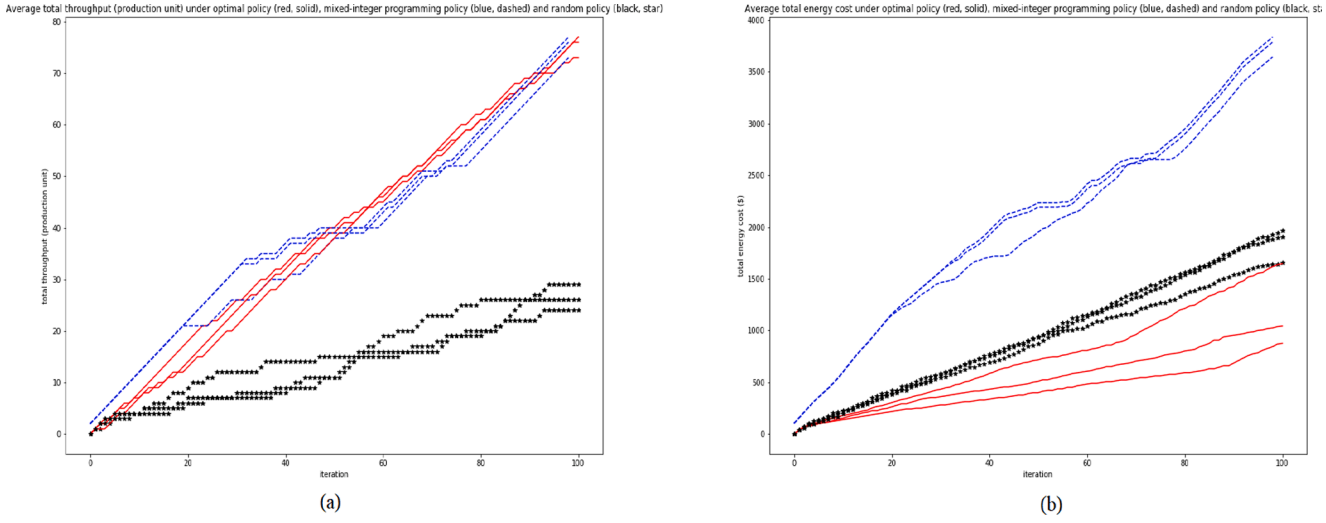


Fig. 6. Left to Right: Average over three experiments the comparison of (a) total throughput in production unit; (b) total energy cost incurred by optimal, routine and random policies: red solid line = optimal policy, blue dashed line = routine strategy via integer programming, black star line = random policy.

microgrid-manufacturing model at a time horizon of 100 time periods, with each time period equals one hour.

5.1.3. Baseline scenarios

The results are compared with two baseline scenarios: The first one runs under a random policy, while the second one is a routine policy. For the *random policy*, the accumulation of total incurred cost $E(\boldsymbol{S},\boldsymbol{A})$ at (16), total energy cost $TF(\boldsymbol{S},\boldsymbol{A})$ at (17) and total production units pr_t in (26) for the optimal policy and random policy are calculated for the system running at a total horizon of 100 time periods (each period = 1 hour).

For the *routine policy*, we consider a routine practice strategy that can be adopted by many industrial practitioners, i.e., the production system and microgrid are controlled or scheduled separately. The production scheduling is generated to minimize total energy consumption without sacrificing target production. This model is briefly introduced as follows. Let x_{it} be the binary decision variable denoting the production schedule of the manufacturing system, i.e., it takes the value of one when machine i is scheduled for production in period t, and zero otherwise. The objective function can be formulated as:

$$\min_{x_{it}} \sum\nolimits_{t\in\mathbf{T}} x_{it}\cdot p_i\cdot\Delta t \tag{56}$$

where $\mathbf{T}$ is the set including all time periods t, p_i is the rated power of machine i, Δt is the time duration of each discretization period. Note that for simplicity, the values of PC_i^{Opr} are used as the rated power without considering the difference between PC_i^{Opr} and PC_i^{Idl}.

Two constraints are formulated as follows:

$$\sum\nolimits_{t\in\mathbf{T}} x_{Nt}\cdot PR_N \geq TA \tag{57}$$

where x_{Nt} is the decision variable for machine N (i.e., the last machine) of the production system. PR_N is the production rate of machine N. TA is the target production count. This constraint shows that the target production should be satisfied. Note that this constraint is based on a simplified assumption that machine breakdown and the resultant blockage/starvation are not considered.

$$0 \leq B_{i(t+1)} = B_{it} + x_{it}\cdot PR_i - x_{(i+1)t}\cdot PR_{i+1} \leq C_i \tag{58}$$

where C_i is the capacity of buffer i. $B_{i(t+1)}$ is the count of work-in-progress parts stored in buffer location i at the beginning of period $t+1$. This constraint shows material flow balance and the work-in-progress part in each buffer location cannot exceed respective bounds.

After solving the aforementioned Integer Programming, the production schedule that can minimize energy consumption without sacrificing production can be obtained. Then, the utilization of microgrids following the empirical rules will be implemented. First, the battery storage system and generator are typically considered backups for emergency situations in practice, and thus are not used in this routine policy. Second, if the renewable sources are available at time period t, they will be first used to satisfy the energy demand of production. If the renewable sources have a higher supply capability than production demand at period t, the remaining part will be sold back to the grid. If the renewable sources have a less supply capability than production demand at period t, the demand gap will be filled by purchasing electricity from the grid. The wind energy has a higher priority to be used than solar source since the cost of wind energy is typically lower than solar energy.

To match the target production unit made by optimal policy, for this routine strategy found by Integer Programming, we set the target output at the time horizon 100, i.e., the target production unit TA in (56) to be equal to 73, which is the total production throughput in units for the optimal policy (the quantity pr_t in (26)) at time horizon 100.

5.2. Result analysis

5.2.1. Random strategy

The system under randomly chosen policy starts with the same initial conditions as the system under optimal policy. The results of 3 experiments are shown in Fig. 6, where red solid lines are for the optimal policy selected by reinforcement learning and black star lines are for the random policy. It is clearly seen from these results that under the optimal policy the manufacturing system tends to produce more throughput with less total cost and similar or less total energy cost (energy demand). More precisely, in one of the experiments, the optimal policy found by reinforcement learning over a time horizon of 100 time periods has an output (the quantity pr_t in (26) of 73, while the randomly selected policy only produces 24. At the same time, the total cost for the optimal policy is $-\$728,293$ and the total cost for the random policy is $-\$236,214$. In terms of energy cost, optimal policy is also about one time less than the random policy, with \$876 for optimal policy and $\$1,656$ for random policy. The two other experiments behave very similarly.

5.2.2. Routing strategy

We found that the total energy cost under the optimal policy found by this integer programming is \$3,642. This has been about four times of our previously announced \$876 for the optimal policy found by reinforcement learning. Two other experiments are made and the results are similar. The results of the evolution of the total energy cost and total production throughput in units as a function of time are shown in Fig. 6, where red solid lines are for the optimal policy selected by reinforcement learning and blue dashed lines are for the routine strategy found by an integer programming. It is clearly seen that the reinforcement learning model selects a policy that incurs less energy cost.

5.2.3. Comparisons

The overall comparison between proposed reinforcement learning model and random policy as well as routine policy is summarized in Table 6, where the last column also compares the average training time (CPU time over 3 experiments) consumed in training for the routine strategy and our proposed reinforcement learning method in the paper.

6. Conclusion

This paper proposes a joint dynamic control model for microgrids and manufacturing systems using MDP to identify an optimal control strategy for both microgrid components and manufacturing system so that the energy cost for production can be minimized without sacrificing production throughput. A novel reinforcement learning algorithm that leverages both TD and DPG algorithms is proposed to solve the joint control of microgrid and manufacturing system towards cost optimality. Experiments for a manufacturing system with an onsite microgrid with renewable sources have been implemented and the results show the effectiveness of the proposed method in addressing the "curse of dimensionality" in dynamic decision-making with high dimensional and complicated state and action spaces.

There are several remarks and future directions that we would like to address here: (1) One can consider real time decision making for emergency situations such as natural disasters that lead to non-availability of external energy supplies from grids (i.e., some new relevant constraints need to be added); (2) Our method could also be extended to other types of manufacturing systems, e.g., flexible manufacturing system, for an extended application scope; (3) The time consumed for the training process in reinforcement learning can be reduced by using reinforcement learning schemes with more exploration, such as ε-greedy [67], Monte-Carlo tree search [79] or upper-confidence-bound method [80].

CRediT authorship contribution statement

Jiaojiao Yang: Data curation, Writing – original draft, Software. **Zeyi Sun:** Conceptualization, Methodology, Writing – original draft. **Wenqing Hu:** Methodology. **Louis Steimeister:** Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Mahieux and A. Oudalov. Microgrids Enter the Mainstream, RenewableEnergyFocus.com, http://www.renewableenergyfocus.com/view/43345/microgrids-enterthe-mainstream/, 2015.
[2] U.S. Department of Energy. Technical Report: How Microgrids Work. http://www.energy.gov/articles/how-microgrids-work, 2014.
[3] Lawrence Berkeley National Laboratory, Technical Report: Microgrid Definitions. https://building-microgrid.lbl.gov/microgrid-definitions, 2016.
[4] Lasseter B. Microgrids [distributed power generation]. In Power Engineering Society Winter Meeting, IEEE 2001;1:146–9.
[5] Lasseter B. Microgrid. In Power Engineering Society Winter Meeting, IEEE 2002;1:305–8.
[6] Hu W, Sun Z, Zhang Y, Li Y. Joint manufacturing and onsite microgrid system control using markov decision process and neural network integrated reinforcement learning. Procedia Manuf 2019;39:1242–9.
[7] Zeng P, Li H, He H, Li S. Dynamic energy management of a microgrid using approximate dynamical programming and deep recurrent neural network learning. IEEE Trans Smart Grid 2019;10(4):4435–45.
[8] Ji Y, Wang J, Xu J, Fang X, Zhang H. Real-time energy management of a microgrid using deep reinforcement learning. Energies 2019;12:2291.
[9] M.A.A. Faruque. RAMP: Impact of rule based aggregator business model for residential microgrid of prosumers including distributed energy resources. In Proceeding of IEEE PES Innovative Smart Grid Technologies Conference (ISGT), 1-6, 2014.
[10] Kriett PO, Salani M. Optimal control of a residential microgrid. Energy 2012;42(1):321–30.
[11] L. Roggia, C. Rech, L. Schuch, J.E. Baggio, H.L. Hey, and J.R. Pinheiro. Design of a sustainable residential microgrid system including PHEV and energy storage device. In Proceedings of the 2011 14th European Conference on Power Electronics and Applications, 1–9, 2011.
[12] F. Ahourai and M.A. Al Faruque. Technical Report: Grid Impact Analysis of a Residential Microgrid under Various EV Penetration Rates in GridLab-D. Center for Embedded Computer Systems, Irvine, CA, 2013.
[13] Igualada L, Corchero C, Cruz-Zambrano M, Heredia FJ. Optimal energy management for a residential microgrid including a vehicle-to-grid system. IEEE Trans Smart Grid 2014;5(4):2163–72.
[14] Hawkes AD, Leach MA. Cost-effective operating strategy for residential micro-combined heat and power. Energy 2007;32(5):711–23.
[15] Tasdighi M, Ghasemi H, Rahimi-Kian A. Residential microgrid scheduling based on smart meters data and temperature dependent thermal load modeling. IEEE Trans Smart Grid 2014;5(1):349–57.
[16] H. Kakigano, Y. Miura, T. Ise, T. Momose, and H. Hayakawa. Fundamental characteristics of DC microgrid for residential houses with cogeneration system in each house. In Proceedings of IEEE Power Energy Society General Meeting-Convers. Del. Elect. Energy 21st Century, 18, 2008.
[17] Olivares DE, Mehrizi-Sani A, Etemadi AH, Canizares CA, Ira-vani R, Kazerani M, et al. Trends in microgrid control. IEEE Trans Smart Grid 2014;5(4):1905–19.
[18] Malysz P, Sirouspour S, Emadi A. An optimal energy storage control strategy for grid-connected microgrids. IEEE Trans Smart Grid 2014;5(4):1785–96.
[19] Parisio A, Rikos E, Glielmo L. A model predictive control approach to microgrid operation optimization. IEEE Trans Control Syst Technol 2014;22(5):1813–27.
[20] Petrollese M, Valverde L, Cocco D, Cau G, Guerra J. Real-time integration of optimal generation scheduling with MPC for the energy management of a renewable hydrogen-based microgrid. Appl Energy 2016;166:96–106.
[21] Craparo E, Karatas M, Singham DI. A robust optimization approach to hybrid microgrid operation using ensemble weather forecasts. Appl Energy 2017;201:135–47.
[22] Su W, Wang J, Roh J. Stochastic energy scheduling in microgrids with intermittent renewable energy resources. IEEE Trans Smart Grid 2014;5(4):1876–83.
[23] Farzan F, Jafari MA, Masiello R, Lu Y. Toward optimal day-ahead scheduling and operation control of microgrids under uncertainty. IEEE Trans Smart Grid 2015;6(2):499–507.
[24] Li Z, Zang C, Zeng P, Yu H. Combined Two-Stage Stochastic Programming and Receding Horizon Control Strategy for Microgrid Energy Management Considering Uncertainty. Energies 2016;9:499.
[25] Mohamed FA, Koivo HN. Online management genetic algorithms of microgrid for residential application. Energy Convers Manage 2012;64:562–8.
[26] M. Ross, R. Hidalgo, C. Abbey, and G. Joós. 2011. Energy storage system scheduling for an isolated microgrid. IET Renewable Power Generation; 5(2): 117-123, 2011.
[27] J. Proano. Microgrid power flow study in grid-connected and islanding modes under different converter control strategies. In Proceedings of 2012 IEEE Power and Energy Society General Meeting, pp. 18, 2012.
[28] H. Yang, F. Wen, and L. Wang. Newton-Raphson on power flow algorithm and Broyden method in the distribution system. In Proceedings of PECon, IEEE 2nd International Power and Energy Conference, pp. 1613-1618, 2008.
[29] J. Driesen, J., and K. Visscher. Virtual synchronous generators. In Proceedings of IEEE Power Energy Society General Meeting Convers. Del. Elect. Energy 21st Century, pp.13, 2008.
[30] Diaz G, Gonzalez-Moran C, Gomez-Aleixandre J, Diez A. Scheduling of droop coefficients for frequency and voltage regulation in isolated microgrids. IEEE Transactions on Power System 2010;25(1):489–96.
[31] Sadegheih A. Optimal design methodologies under the carbon emission trading program using MIP, GA, SA. TS Renewable and Sustainable Energy Reviews 2011;15(1):504–13.
[32] Chaouachi A, Kamel RM, Andoulsi R, Nagasaka K. Multiobjective intelligent energy management for a microgrid. IEEE Trans Ind Electron 2013;60:1688–99.
[33] Duflou JR, Sutherland JW, Dornfeld D, Herrmann C, Jeswiet J, Kara S, et al. Towards energy and resource efficient manufacturing: A processes and systems approach. CIRP Annals-Manufacturing Technology 2012;61(2):587–609.
[34] U.s.. Department of Energy. Annual Energy Review 2009;ftp://ftp.eia.doe.gov/multifuel/038409.pdf:2010.
[35] Ma S, Zhang Y, Liu Y, Yang H, Lv J, Ren S. Data-driven sustainable intelligent manufacturing based on demand response for energy-intensive industries. J Cleaner Prod 2020;274:123155.
[36] Zhang Y, Ma S, Yang H, Lv J, Liu Y. A big data driven analytical framework for energy-intensive manufacturing industries. J Cleaner Prod 2018;197:52–72.
[37] Lv J, Tang R, Jia S. Therblig-based energy supply modeling of computer numerical control machine tools. J Cleaner Prod 2014;65:168–77.
[38] Yun L, Ma S, Li L, Liu Y. CPS-enabled and knowledge-aided demand response strategy for sustainable manufacturing. Adv Eng Inf 2022;52:101534.
[39] Yun L, Li L, Ma S. Demand response for manufacturing systems considering the implications of fast-charging battery powered material handling equipment. Appl Energy 2022;310:118550.
[40] Sun Z, Li L, Fernandez M, Wang J. Inventory control for peak electricity demand reduction of manufacturing systems considering the tradeoff between production loss and energy savings. J Cleaner Prod 2014;82:84–93.
[41] Sun Z, Dababneh F, Li L. Joint energy, maintenance, and throughput modeling for sustainable manufacturing systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2020;50(6):2101–12.
[42] F. Katiraei, C. Abbey, S. Tang, and M. Gauthier. Planned islanding on rural feeders—utility perspective. In Proceedings of Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, IEEE, pp.1-6, 2008.
[43] J. Turkewitz. Unemployment Deepens Storm's Loss as Businesses Stay Closed. http://www.nytimes.com/2012/12/28/nyregion/unemploymentdeepens-the-loss-from-hurricane-sandy.html? r=0, 2012.
[44] M.M. Islam, Z. Sun, and X. Yao. Simulation-based investigation for the application of microgrid with renewable sources in manufacturing systems towards sustainability. In ASEM 2016 International Annual Conference, Charlotte, NC, USA. American Society for Engineering Management, 2016.
[45] Islam MM, Sun Z. Onsite generation system sizing for manufacturing plant considering renewable sources towards sustainability. Sustainable Energy Technol Assess 2019;32:1–18.
[46] Zhong X, Islam MM, Xiong H, Sun Z. Design the capacity of onsite generation system with renewable sources for manufacturing plant. In Procedia Computer Science, Complex Adaptive Systems Conference, Chicago, IL, USA 2017;114:433–40.
[47] Santana-Viera V, Jimenez J, Jin T, Espiritu J. Implementing factory demand response via onsite renewable energy: A design-of-experiment approach. Int J Prod Res 2015;53(23):7034–48.
[48] T.J. Harper. A Novel Microgrid Demand-Side Management System for Manufacturing Facilities. Master thesis, 2014. Purdue University. http://docs.lib.purdue.edu/techmasters/85/.
[49] T.J. Harper, W.J. Hutzel, A. Kulatunga, J.C. Foreman, and A.L. Adams. Microgrids for Improving Manufacturing Energy Efficiency. In Proceedings of International High Performance Buildings Conference, 2014.
[50] Golari M, Fan N, Jin T. Multistage Stochastic Optimization for Production-Inventory Planning with Intermittent Renewable Energy. Production and Operations Management 2017;26(3):409–25.
[51] Reducing energy use by the pulp and paper industry will require greater recycling and waste heat recovery, IEA, https://www.iea.org/fuels-and-technologies/pulp-paper.
[52] Pandey AK, Prakash R. Energy Conservation Opportunities in Pulp & Paper Industry. Open Journal of Energy Efficiency 2018;7(4).
[53] Ashok S. Peak-Load Management in Steel Plants. Appl Energy 2006;83(5):413–24.
[54] Ashok S, Banerjee R. An Optimization Mode for Industrial Load Management. IEEE Trans Power Syst 2001;16(4):879–84.
[55] Ma S, Zhang Y, Lv J, Yang H, Wu J. Energy-cyber-physical system enabled management for energy-intensive manufacturing industries. J Cleaner Prod 2019;226:892–903.
[56] Ma S, Zhang Y, Lv J, Ge Y, Yang H, Li L. Big data driven predictive production planning for energy-intensive manufacturing industries. Energy 2020;211:118320.
[57] Ma S, Zhang Y, Ren S, Yang H, Zhu Z. A case-practice-theory-based method of implementing energy management in a manufacturing factory. Int J Comput Integr Manuf 2021;34:829–43.
[58] Zhang H, Sun W, Li W, Wang Y. Physical and chemical characterization of fugitive particulate matter emissions of the iron and steel industry. Atmos Pollut Res 2022;13:101272.
[59] W. Sun, Q. Wang, Y, Zhou, and J. Wu. Material and energy flows of the iron and steel industry: Status quo, challenges and perspectives, Applied Energy, 268: 114946, 2020.

[60] Zhang H, Sun W, Li W, Ma G. A carbon flow tracing and carbon accounting method for exploring CO2 emissions of the iron and steel industry: An integrated material–energy–carbon hub. Appl Energy 2020;309:118485.
[61] Meirina C, Levchuk YN, Levchuk GM, Pattipati KR. A markov decision problem approach to goal attainment. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 2008;38(1):116–32.
[62] Doshi P, Qu X, Goodie AS, Young DL. Modeling human recursive reasoning using empirically informed interactive partially observable markov decision processes. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 2012;42(6):1529–42.
[63] E. Kuznetsova, Y. F. Li, C. Ruiz, E. Zio, G. Ault, and K. Bell. Reinforcement learning for microgrid energy management. Proc. of the 35th International Conference on Uncertainty in Artificial Intelligence (UAI), Tel Aviv, Israel, 59:133-146, 2013.
[64] Dallery Y. On modeling failure and repair times in stochastic models of manufacturing systems using generalized exponential distributions. Queueing Systems 1994;15(1–4):199–209.
[65] Wang Y, Li L. A novel modeling method for both steadystate and transient analyses of serial bernoulli production systems. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2014;45(1):97–108.
[66] J. Li and S.M Meerkov. Production systems engineering. Springer Science & Business Media.
[67] R.S. Sutton and A.G. Barto. Reinforcement Learning: An Introduction. Second Edition, in progress, complete draft online. MIT Press, November 5, 2017.
[68] Konda VR, Tsitsiklis JN. Actor-Critic Algorithms. Neural Information and Processing Systems 2000.
[69] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. Neural Information and Processing Systems (NIPS), 2013.
[70] D.P. Bertsekas and J.N. Tsitsiklis. Neuro–Dynamic Programming. Athena Scientific, 1996.
[71] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. ICML (International Conference on Machine Learning), 2014.
[72] Parikh N, Boyd S. Proximal algorithms. Foundations and Trends in Optimization 2013;1(3):123–231.
[73] Janner M, Fu J, Zhang M, Levine S. When to Trust Your Model: Model-Based Policy Optimization. Neural Information and Processing Systems 2019.
[74] Github-Hu, 2020, Source code for our paper: https://github.com/huwenqing0606/rlmanufacturing.
[75] Li L, Sun Z. Dynamic energy control for energy efficiency improvement of sustainable manufacturing systems using Markov Decision Process. IEEE Transactions on Systems, Man, and Cybernetics: Systems 2013;43:1195–205.
[76] Solar Energy Local: Solar energy data and resources in the us. https://solarenergylocal.com/.
[77] State Climatologist Office for Illinois. http://www.isws.illinois.edu/atmos/statecli/wind/wind.htm.
[78] D.P. Kingma and J. Ba. Adam: A method for Stochastic Optimization. ICLR, arXiv: 1412.6980[cs.LG], 2015.
[79] Browne C, Powley E, Whitehouse D, Lucas S, Cowling PI, Rohlfshagen P, et al. A Survey of Monte Carlo Tree Search Methods. IEEE Trans Comput Intell AI Games 2012;4(1):1–43.
[80] Auer P. Using Confidence Bounds for Exploitation-Exploration Tradeoffs. Journal of Machine Learning Research 2002;3:397–422.