Renewable and Sustainable Energy Reviews 137 (2021) 110618

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: http://www.elsevier.com/locate/rser

Applications of reinforcement learning in energy systems

A.T.D. Perera [a,*], Parameswaran Kamalaruban [b]

[a] Urban Energy Systems Laboratory, EMPA, Überland Str. 129, 8600, Dübendorf, Switzerland
[b] The Alan Turing Institute, London, United Kingdom

ARTICLE INFO

Keywords:
Energy systems
Reinforcement learning
Renewable energy
Building energy
Machine learning

ABSTRACT

Energy systems undergo major transitions to facilitate the large-scale penetration of renewable energy technologies and improve efficiencies, leading to the integration of many sectors into the energy system domain. As the complexities in this domain increase, it becomes challenging to control energy flows using existing techniques based on physical models. Moreover, although data-driven models, such as reinforcement learning (RL), have gained considerable attention in many fields, a direct shift into RL is not feasible in the energy domain irrespective of the ongoing complexities. To this end, a top-down approach is used to understand this behavior by reviewing the current state of the art.

We classified RL papers in the literature into seven categories based on their area of application. Subsequently, publications under each category were further examined relative to problem diversity, RL technique employed, performance improvement (compared with other white and gray box models), verification, and reproducibility; many of the articles reported a 10–20% performance improvement with the use of RL. In most studies, however, deep learning techniques and state-of-the-art actor-critic methods (e.g., twin delayed deep deterministic policy gradient and soft actor-critic) were not applied. This has remarkably hindered performance improvements and problems related to complex energy flows have not been considered. Approximately half of the publications reported the use of Q-learning. Furthermore, despite the availability of historical data in the energy system domain, batch RL algorithms have not been exploited. Emerging multi-agent RL applications may be considered as a positive development that can enable the management of complex interactions among multiple parties. Most studies lack proper benchmarking compared to model-based approaches or gray-box models, and a majority cover energy dispatch problems and building energy management. Although RL can adequately solve problems that are considerably integrated in several sectors, only a limited number of publications have discussed its broad application. The present study clearly demonstrates that even without the full utilization of RL capacity, this technique has a considerable potential in resolving the continuously increasing complexity within the energy system domain.

1. Introduction

With the escalating accumulation of CO_2 emissions in the atmosphere, the frequent occurrence of extreme climate events, and the rapid increase in global population particularly in urban areas, significant changes in the energy sector are necessary [1]. It is anticipated that energy sustainability and energy efficiency improvement will perform vital functions in the urban sector where the integration of sustainable energy technologies are necessary [2,3]. In addition, the energy nexus, such as between water, agriculture, and transportation, should be considered as these elements tend to improve the sustainability of several sectors while minimizing greenhouse gas emissions [4]. However, the introduction of these changes into energy systems is an exigent task when both demand and generation are taken into account [5].

1.1. Problems on growing energy system complexity

As the complexities in the energy sector increase, it becomes more difficult to optimally control energy systems. For example, centralized generation is gradually moving into distributed energy systems, replacing fossil fuel-based dispatchable energy sources from renewable energy technologies [6]. The inclusion of renewable energy technologies (e.g., solar photovoltaics (PV), solar thermal, and wind) makes the

* Corresponding author.
E-mail address: dasun.perera@empa.ch (A.T.D. Perera).

https://doi.org/10.1016/j.rser.2020.110618
Received 1 May 2020; Received in revised form 20 October 2020; Accepted 21 November 2020
Available online 9 December 2020

operation of distributed energy systems more problematic because of the intermittent nature of these sources. Energy storage and dispatchable energy technologies, such as combined heat and power (CHP) generators, are necessary because of the short and long-term changes (stochastic nature) in renewable energy potential and energy demand [7]. It is difficult to integrate these components into a single system because of the intermittent nature of energy potentials, variations in demand, and complexities in energy conversion processes [8,9]. Similarly, the complexity of the operation increase when introducing energy storage to work in harmony with the internal combustion engines in automobiles [10]. In the energy sector, these changes become increasingly common because of the demand for CO_2 emission reduction.

Energy transition introduces problems that are well beyond the boundaries of energy systems; for example, the energy nexus between transportation, agriculture, waste management, and buildings also requires considering the interaction among these sectors. This typically necessitates co-simulation platforms that lead to bulky models [11], which are difficult to employ for control purposes. Furthermore, existing models focus on presenting the physical interactions among the different sectors, and usually fail to take into account cyber interactions [12]. Considering both cyber and physical interactions for control purposes using existing model-based approaches is another problem [13]. In addition to the limitations caused by the increasing complexity of energy systems, particular attention should be devoted to uncertainty and security management [14]. Uncertainties, such as climate change, energy market variations, and improvements in energy technologies, are considered using bulky physical models that demand extensive computational time especially for control purposes [15–17]. Therefore, the multi-dimensional impact of certain uncertainties are often neglected, thereby making it difficult to assess the effect of certain critical phenomena (e.g., climate change) on the energy sector [18]. In view of these limitations in the present state-of-the-art techniques, significant changes in the modeling methods employed in the design and operation sectors are necessary. In conclusion, with the increasing complexity in energy systems, cyber-physical interactions, uncertainties, and security challenges, a paradigm shift in the present state-of-the-art methodologies for energy system control is required [5].

1.2. Emergence of data science

The application of machine learning techniques have become a main research focus irrespective of the research domain especially with the emergence of deep learning [19]. The number of research publications on machine learning has rapidly increased, and machine learning methods have gradually attracted the attention of researchers in the energy sector for managing the aforementioned complexities because of the model-free approach of these techniques [20]. Machine learning employ a data-driven methodology that can support energy experts in considering complex planning problems at the urban, regional, and national scales [20]. The main branches of machine learning (i.e., supervised, semi-supervised, unsupervised, and reinforcement) are already well-established in the energy domain [21,22] and can be facilely understood by referring to energy-related publications that elaborate on machine-learning techniques. In shifting into the energy system domain, machine-learning techniques are employed in all the major steps of energy system design process. The number of publications that discuss machine learning techniques covering topics ranging from renewable energy forecasting to development of complex surrogate models that can be used for energy system design has rapidly increased (Fig. 1).

1.3. Emergence of reinforcement learning

Reinforcement learning (RL), a branch of machine learning, incorporates human-level control [23,24], which has attracted attention in several fields. Although it is observed that the application of RL in the energy domain has gained considerable interest, there is a reasonable time lag between such an application and the publication of papers that present machine learning (including RL). A lag is also evident when RL publications are compared with those that focus on model predictive control (MPC) methods (Fig. 2). Since 2008, there has been a certain gap between the number of publications that featured MPC and RL (irrespective of the domain); it was only in 2018 when the compensation for this deficiency was introduced to a moderate extent (Fig. 2). However, in moving into the energy domain, a gap, which tends to further increase, is observed. This is unexpected particularly with the increasing complexity of energy systems, uncertainties, and security problems, which are difficult to control entirely using model-based approaches. Accordingly, it is important to implement a more holistic analysis of the present state-of-the-art applications of RL in the energy sector to identify the root causes of the gap. This requires a more thorough assessment besides a mere paper review. Accordingly, this study uses a top-down approach to review the present state-of-the-art applications of RL.

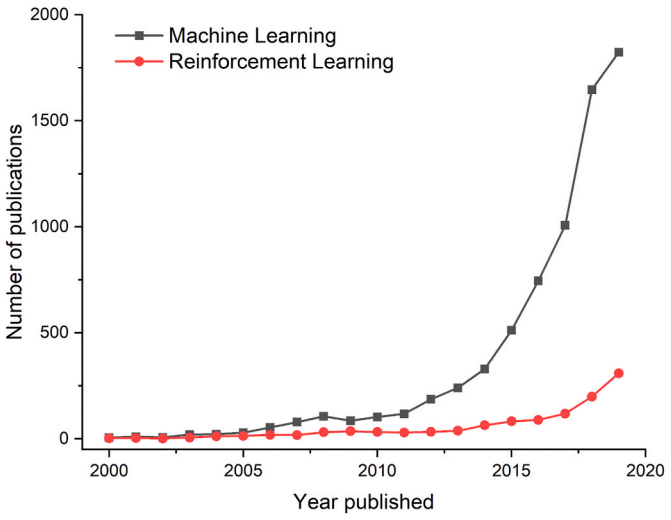

Fig. 1. Title, abstract, and keyword search on machine learning and reinforcement learning including energy systems (2000–2019); extracted November 28, 2019.

1.4. Objectives and methodology

In literature, several papers present a comprehensive overview of the state-of-the-art methods. Cheng and Yu [25] extensively reviewed the machine learning methods implemented in the energy and electric power system domains that mainly include many aspects in the energy sector and RL. Their paper was intended to provide a general overview of the state of the art instead of simply reviewing each article since the time machine learning techniques started to be extensively used in the energy domain. Han et al. [26] focused only on RL and the control of occupant comfort in buildings. Vázquez-Canteli and Nagy [27] discussed RL applications in demand response as well as building control problems related to demand, generation, and energy management. These studies employed a paper-by-paper approach and presented the previous studies that were conducted within a definite scope. Recently [28], reviewed the recent progress in building energy management systems using the same approach.

All these studies provides a comprehensive overview about the use of RL in building energy systems. A paper by paper review is presented on this regard where the specific application as well as the RL method are classified in a comprehensive manner. The number of publications related to RL grow at a rapid speed. Therefore, it is hard to track all the papers especially for someone starting new which is quite common within the machine learning concerning areas such as computer vision,

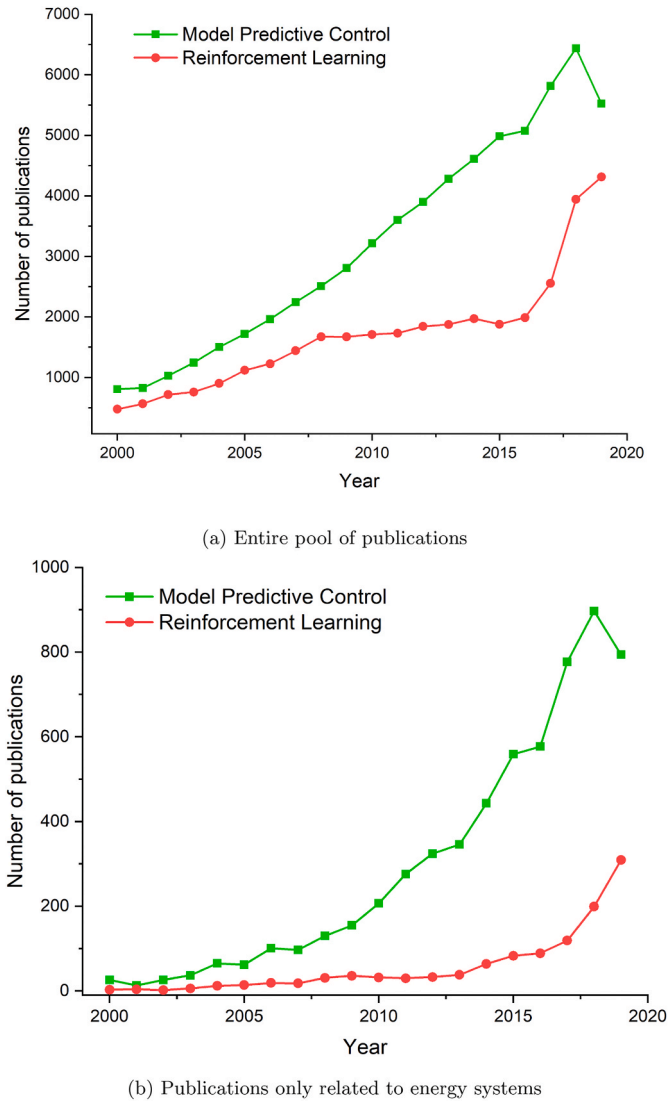

Fig. 2. Number of publications obtained using title, abstract, and keyword search on model predictive control and reinforcement learning: (a) entire pool of publications and (b) publications related only to energy systems (2000–2019); extracted November 28, 2019.

supervised learning etc. Where a large pool of papers appear each year. Although energy transition introduces several energy management problems, which cannot be resolved by the mere use of classical control theory-based approaches, most of these problems having many similarities because they are related to energy flow management. Hence, it is possible to develop a common knowledge base. However, this is an extremely exigent task because the collection and organization of relevant literature require considerable effort. It is also more difficult to present following a paper-by-paper format similar to that in Ref. [26, 27]. Accordingly, in this study, the research papers related to RL applications in the energy system domain are classified into several categories and subsequently extended by considering the crosslinks among these classes. Section 2 provides a comprehensive overview about the control/operation problems where RL has been used leading into a classification of these applications into six main classes. Being similar to the applications, reinforcement methods can be classified into several classes which makes it easy to understand the use of RL algorithm. Section 3 presents a classification of RL techniques used in the present state of the art with a comprehensive theoretical background on each method. A detailed cross comparison among problem classes and RL techniques are performed in Section 4 being focused on:

- The complexity of the control problem considered within the domain
  - Handling non-linearity
  - Use of approximation models/data driven models
  - Expansion of decision and objective space variables
- Similarity among the different classes of problems
- Verification of results and reproducibility of approaches
- Computational burden for problem class

Based on Section 4, Section 5 is devoted to discuss about the future perspectives on RL applications in the energy system sector, particularly on extending the boundaries of RL problems related to sector coupling, linking control problems with multi-resolution time steps, and shifting focus from control to coupled control and design problems, are discussed. Finally, Section 6 presents the conclusions of the study.

2. Broader applications of RL

The integration of variable renewable energy technologies introduces problems into the energy system domain from the perspectives of control, stability, and security. It is therefore important to understand the advantages afforded by the use of RL in dealing with these problems. In Section 2.1, the benchmarking of RL with other existing techniques in order to better comprehend these advantages is presented. As shown in Fig. 3, the peak energy generated by renewable energy technologies may not satisfy demand, making it essential to depend on other energy technologies or energy storage. The stochastic nature of demand and generation performs a vital function when managing energy flows within the system. This leads to a set of operation optimization problems in scheduling generation (commonly known as energy dispatch problem), energy system operation within buildings, device control (e.g., PV panels), and market interaction, which are deemed as the main applications of RL; thus, in this domain, RL application is not limited to the dispatch problem. Section 2.2 elaborates and classifies the various RL applications in the energy system domain.

2.1. Incorporation of RL in energy system domain

Several different approaches are employed in the state of the art to control energy systems. These can be categorized into three classes: white box, gray box, and data-driven (black box) models [29]. The white box models, which are also known as model-based control strategies, apply physical principles to represent the relationship between model inputs and outputs during the control process; the model predictive control (MPC) strategy discussed in the Introduction section is classified as a white box model. Data-driven models, also known as black box control methods, use the knowledge derived by processing online or offline data instead of depending on the explicit or implicit information

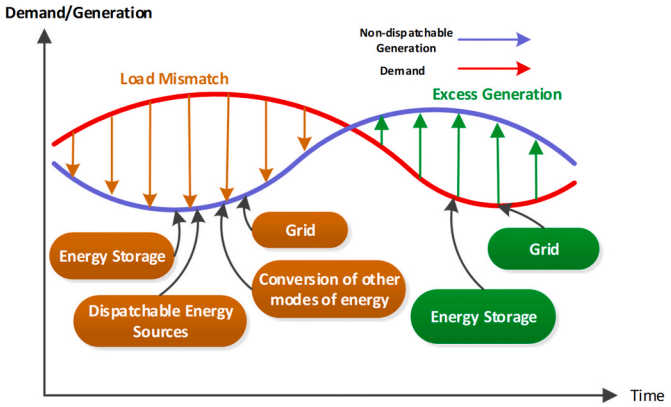

Fig. 3. Requirements for optimal control to avoid mismatch between demand and generation.

of the mathematical model; RL belongs to the data-driven category. Gray box models are those that are between white and black box models; models based on fuzzy logic are among those classified under this topic. In this section, these approaches are differentiated to aid readers gain a better understanding of the techniques rather than presenting a broad categorization of existing modeling tools in the energy domain.

Consider a simple energy system with renewable energy technologies and energy storage (Fig. 4), i.e., the previously explained dispatch problem. The energy system is connected to the grid and caters to the electricity demand in a neighborhood. When sufficient renewable energy is not generated to satisfy the energy demand, either a battery bank or the grid can be used to compensate for the mismatch. The choice depends on several factors, such as the current price of electricity in the grid, price forecast of electricity in the grid for the time horizon, renewable energy generation forecast for the time horizon, and demand forecast for the time horizon. The white box approach uses either dynamic programming or MPC technique to select the appropriate control decision. This approach uses a detailed model to represent the energy and cash flow within the energy system that are subsequently linked to an optimization algorithm to derive the optimal states for the control horizon. However, the uncertainties in the forecast may also perform a vital role in such instances; in this case, stochastic MPC and stochastic dynamic programming techniques are used. Model dependencies and convexity guarantee make it considerably problematic to extend such an approach to more complex energy systems [30]. In such instances, the gray box models use approximation methods, such as fuzzy logic, to assist in achieving dispatch decisions [31]. Although fuzzy rules are typically defined, the pool of fuzzy decisions significantly increases with the complexity of the energy system, making optimization difficult [32]. In this regard, RL takes a different approach that is solely based on a data-driven model by employing an agent to participate in the process. The agent makes decisions based on the data-driven model and accepts inputs from the surroundings. The data-driven model that makes control decisions is fine-tuned by participating in the decision-making process and maximizing its reward. There are a number of different methodologies under the broad scope of RL that may be utilized to formulate and train the data-driven model; these are discussed in detail in Section 3. The data-driven model is effective in managing model uncertainties and complexities, which are difficult to achieve using white box models.

2.2. Classification of RL problems in energy systems

The changes in outdoor conditions (e.g., fluctuations in temperature, solar irradiation, and wind speed) influence generation and demand. Accordingly, the entire energy system and each device should maximize its performance (e.g., use of maximum power point tracking for wind turbines and PV panels). Moreover, a number of other factors, such as occupancy and equipment usage pattern, are expected to introduce uncertainties into the demand side. Therefore, considering a standalone operation, the mere matching of demand and generation is already a complex problem. Numerous publications regard demand and generation as two distinct control problems, as shown in Fig. 3. The optimal operation of heating, ventilation, and air conditioning (HVAC) systems that considers changes in the environment is typically regarded as a demand side problem, which neglects uncertainties in power generation. There are a number of recent publications pertaining to vehicles and grids, which well fit the demand and generation sides. In addition, there are numerous instances when the uncertainties in both demand and generation sides are concurrently discussed. Accordingly, it is difficult to define the boundaries of the foregoing problems. On the other hand, specific problems, such as maximum power point tracking (MPPT) of wind-turbines and solar panels, can be defined without infringing on other problems.

By conducting a keyword search in Scopus, two groups of

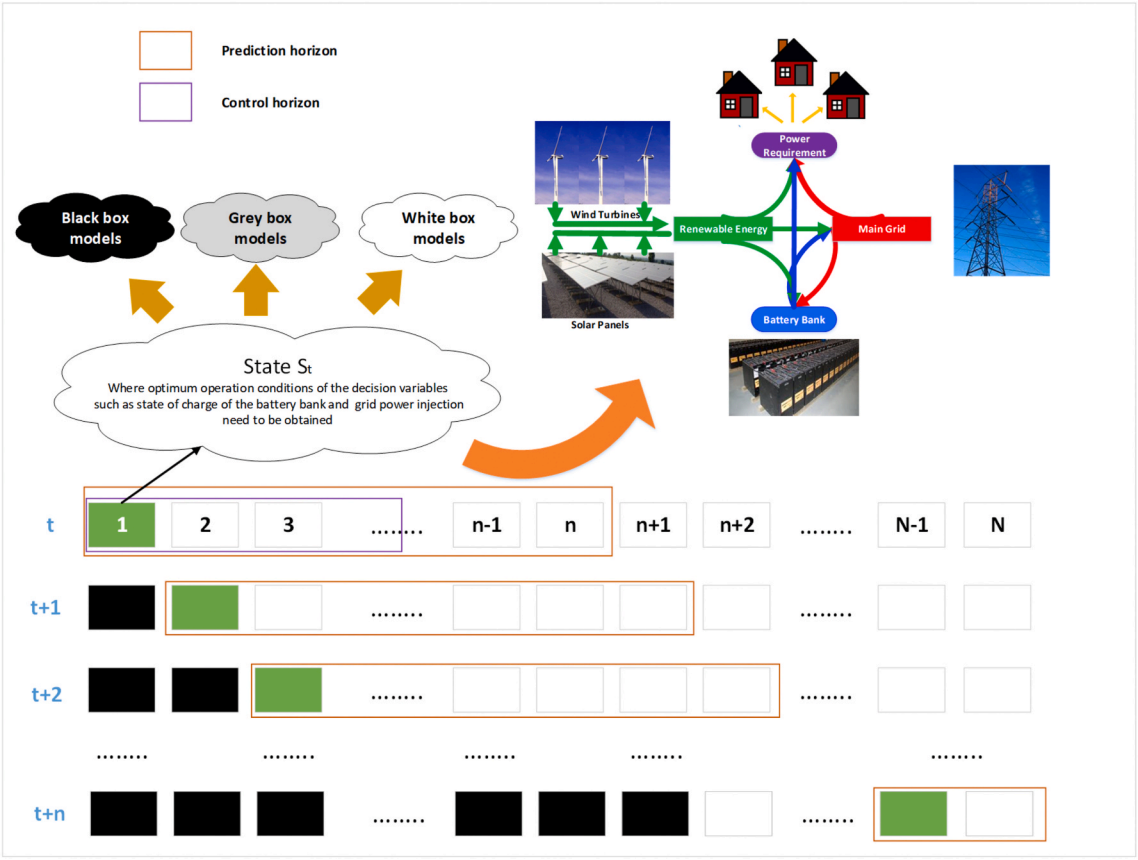

Fig. 4. Graphical overview of dispatch problem.

publications are found: 1) those that deal with specific problems (SPs) and 2) those that deal with integrated problems (IPs). The SP category focuses on a well-defined and extremely confined domain to delve deep into and resolve typically broader IPs. It is extremely exigent to provide an exact definition for each group as there are many gray areas. The SPs are further classified into the following six groups considering the application domain:

- Building energy management system (BEMS)
- Dispatch
- Vehicle energy systems
- Energy devices
- Grid
- Energy markets

The BEMS is a common framework in the energy system domain for implementing optimal control strategies, such as those for HVAC, lighting, and blinds, related to the thermal inertia of buildings, weather uncertainties, and occupant behaviors. Within the BEMS, catering heating, air-conditioning, ventilation demands of the building, job scheduling, optimal control building elements (such as window blinds), and maintaining indoor air quality are often considered. The main focus of BEMS is to either increase the comfort level and energy efficiency or minimize the cost. The scope of the problem notably extends when moving from BEMS to the dispatch problem. The main focus of the dispatch problem is delivering electricity, heat, and cooling demand by optimally using energy storage, renewable energy technologies, and dispatchable energy sources. Some studies take price signals from the grid to determine the optimal dispatch strategy for the energy system. Although it is not common, there are instances that the energy system is expected to cater to diverse applications such as desalination besides being limited to heating, cooling, and electricity. Often cost minimization is considered as the main objective in the dispatch problem. However, minimizing emissions is gradually getting popular due to environmental concerns. It is difficult to define the exact boundary between dispatch problem and BEMS because many publications present problems that can be categorized between these two; hence, IP, BEMS, and dispatch are merely defined to classify the publications that discuss the elements of BEMS and dispatch problems. In terms of RL applications in the transportation sector (within the energy system domain), the control problems can be classified into two: optimal charging-discharging using grid electricity (Vehicle to Grid (V2G)) and energy management problem within the vehicle (vehicle energy system). Although these two are closely interlinked, there is no publication that discusses this aspect together; hence, they are treated separately. A vehicle energy system is regarded as an SP that does not maintain any links with BEMS or dispatch problems. In contrast, the V2G problem is usually associated with BEMS, energy markets, or dispatch problems that are related to the optimal time slots for charging vehicles based on changes in the grid electricity price. Accordingly, the problem is typically extended considering the optimal time slots to charge vehicles, electricity price in the grid, and renewable energy generation that interlinks power demand and generation. Therefore, two IPs are introduced, V2G-dispatch and V2G-BEMS, to consider the interactions with the transportation sector.

The energy system operation (dispatch decisions) is influenced by a number of factors, such as demand, grid conditions, energy market, and changes in the performance of energy system components. The use of RL for grid control involves a broad field of research, the transient stability of the grid, and n-1 security; voltage and frequency regulations, optimal power flow, etc. are considered in this context. In the present study, a detailed description of grid control using RL is not considered, and the scope is only limited to the impact of the grid on the energy system. Promising methods for operating the energy systems while actively participating in the energy markets are evaluated. The participation of energy systems into the day ahead, balancing, and real-time markets are considered in this context. Similar to vehicles, the impacts of energy markets on the BEMS and dispatch problem are separately taken into account, although RL applications are not significant in these areas. Finally, the energy device classification focuses on the optimal control of energy system components (devices), performing an important function to improve the energy system efficiency. In this subset, the focus is set on the maximum power point tracking (MPPT) of wind turbines and solar panels.

3. RL methods and applications in different problem classes

RL is a branch of machine learning that mainly focuses on sequential decision-making that takes into account uncertainties. The recent advances in deep RL have achieved remarkable performance in games [33,34], continuous control [35], and robotics [36]. RL can also be defined as the problem of learning how to act optimally in an environment through experience. In this regard, an RL agent must interact with its environment and learn how to maximize certain cumulative rewards over time. RL has made a reasonable progress during the recent past. This section provides a present state of the art methodologies used within the RL community and used for energy system operation. A reader who is already familiar with RL methods or interested in only RL applications in energy systems can safely skip this section. In this section, Section 3.1 provides the necessary background on RL; Section 3.2 discusses various RL methods used in practice; Section 3.3 summarizes the implementation packages of deep RL algorithms.

3.1. Background

Formally, RL is a game between an agent and an environment. The environment is represented by a Markov Decision Process (MDP) $\mathscr{M} := (\mathscr{S}, \mathscr{A}, T, \gamma, P_0, R)$, where the state and action spaces are denoted by $\mathscr{S}$ and $\mathscr{A}$ respectively. $T : \mathscr{S} \times \mathscr{S} \times \mathscr{A} \to [0,1]$ captures the state transition dynamics, *i.e.*, $T(s'|s,a)$ denotes the probability of landing in state s' by taking action a from state s. Here γ is the discounting factor, $P_0 : \mathscr{S} \to [0,1]$ is the initial distribution over states $\mathscr{S}$, and $R : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ is the reward. RL agent's behavior is fully captured by a policy $\pi : \mathscr{S} \to \mathscr{A}$, which maps states to action. The set Π denotes the set of all stationary policies. An optimal policy intuitively maximizes the overall cumulative (discounted) reward.

The game between the agent and the environment is given as follows:

- At time step $t = 0$: $S_0 \sim P_0(\cdot)$
- At each time step $t = 0, 1, 2, \ldots$:
  - agent observes the environment's state $S_t \in \mathscr{S}$
  - agent chooses an action $A_t = \pi(S_t) \in \mathscr{A}$
  - agent receives a reward $R_{t+1} = R(S_t, A_t)$
  - agent finds itself in a new state $S_{t+1} \sim T(\cdot|S_t, A_t)$

The above game is graphically illustrated in Fig. 5. An agent executing a policy $\pi : \mathscr{S} \to \mathscr{A}$ in the environment $\mathscr{M}$ obtains a random cumulative return $Z = \sum_{t=1}^{\infty} \gamma^{t-1} r_t$, where $r_t = R(S_t, A_t)$. A typical objective of RL is to find an optimal policy, π^*, which maximizes the discounted sum of rewards (1),

$$\pi^* = \arg\max_{\pi : \mathscr{S} \to \mathscr{A}} \mathbb{E}_{\pi} \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \right]. \tag{1}$$

Value Functions: Value functions are estimations of the expected return over a certain time horizon given the current state $s \in \mathscr{S}$ and are often used to construct the optimal policy. In particular, the state-value function V^{π} of a given policy π is defined as (2)

$$V^{\pi}(s) := \mathbb{E}_{\pi} \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| S_0 = s \right], \tag{2}$$

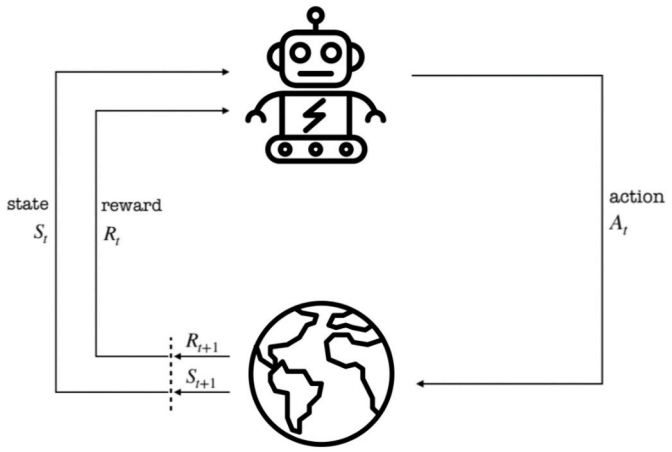

Fig. 5. Interaction between the agent and the environment (source: adapted from Ref. [37]).

which stands for the expected return starting from state $s \in \mathscr{S}$. The optimal value function V^* corresponds to the value function of an optimal policy π^*, i.e., (3):

$$V^*(s) := \max_{\pi:\mathscr{S}\to\mathscr{A}} V^{\pi}(s). \tag{3}$$

Once V^* is available, then the optimal policy can be recovered by picking an action a that is greedy with respect to V^*, i.e., (4):

$$\pi^*(s) = \arg\max_{a\in\mathscr{A}} \mathbb{E}_{s'\sim P(\cdot|s,a)}[R(s,a) + \gamma V^*(s')]. \tag{4}$$

In RL setup, the model information such as transition probabilities and reward distributions are usually unavailable, thus it is not easy to directly obtain the optimal policy from state-value function.

Instead, the action value function or Q-function has been considered, which is defined as (5)

$$Q^{\pi}(s,a) := \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, S_0 = s, A_0 = a\right], \tag{5}$$

and the corresponding optimal Q-function is defined as (6)

$$Q^*(s,a) := \max_{\pi:\mathscr{S}\to\mathscr{A}} Q^{\pi}(s,a). \tag{6}$$

Observe that once Q^* is available, then the optimal policy can be retrieved by $\pi^*(s) = \arg\max_{a\in\mathscr{A}} Q^*(s,a)$, which does not require the model information in contrast to the state-value function case.

Dynamic Programming: When the MDP $\mathscr{M}$ is known, finding π^* is called a *planning* problem, and it can be solved efficiently via dynamic programming (DP) algorithms. The value functions for given π can be computed or estimated by solving the Bellman equations (7) and (8) [38–40]:

$$V^{\pi}(s) = \mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}[R(s,\pi(s)) + \gamma V^{\pi}(s')], \tag{7}$$

$$Q^{\pi}(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}[R(s,a) + \gamma Q^{\pi}(s',\pi(s'))], \tag{8}$$

which are derived from the Markov property [40] of the MDP and the definition of the value function. This step is often called the policy evaluation. Once a policy π is evaluated, then an improved policy can be obtained by the greedy policy, $\pi'(s) = \arg\max_{a\in\mathscr{A}} \mathbb{E}_{s'\sim P(\cdot|s,a)}[R(s,a) + \gamma V^{\pi}(s')]$ or $\pi'(s) = \arg\max_{a\in\mathscr{A}} Q^{\pi}(s,a)$, which is called the policy improvement. Various dynamic programming algorithms, such as the policy and value iterations, are based on various combinations of alternative iterations of the policy evaluation and improvement [39].

3.2. RL methods

RL methodologies can be broadly classified into several classes. Therefore, it is challenging to come up with a comprehensive classification. In this study, we provide a broad classification of RL methodologies by categorizing them into three sets: Value-based, Policy-based, and Model-based (cf. Fig. 6). We can further group the model-based RL methods into the following three groups: (i) value-based, e.g., Dyna-Q algorithm [42], Deep Dyna-Q algorithm [43], and Value-Aware Model Learning (VAML) [44], (ii) policy-based, e.g., Model-Based Policy Gradient (MBPG) [45], ME-TRPO, SLBO, and Policy-Aware Model Learning (PAML) [46], and (iii) actor-critic, e.g., Model-Based Actor--Critic (MBAC) [47], Model-Augmented Actor-Critic (MAAC) [48], and Dyna-DDPG [49]. Some methods belong to several sets. For example, Deterministic policy gradient (DPG) and Deep Deterministic policy gradient (DDPG) belong to both value and policy sets. Within this broad classification, this section is divided into five classes, namely, Value-based, Policy-based, Actor-critic, Model-based, and Batch RL. RL methods can also be categorized into on-policy and off-policy methods. On-policy methods estimate the value of a policy that is used for control. In contrast, off-policy methods evaluate a policy (estimation policy) different from that used to generate behavior (behavior policy).

3.2.1. Value based methods

Value-based RL methods aim to learn the state or action-value function and then to select actions accordingly. SARSA [50] and Q-learning [51] are the two key algorithms in this category.

Temporal-Difference (TD) Learning: TD learning is the most commonly used policy evaluation algorithm. TD learning algorithm estimates the state value function V^{π} of a given policy π iteratively. The update rule for TD learning is derived from the squared-Bellman error and is given by (9) [37]

$$V_{k+1}(s_k) = V_k(s_k) + \alpha_k(R(s_k,\pi(s_k)) + \gamma V_k(s_{k+1}) - V_k(s_k)), \tag{9}$$

where $s_k \sim d^{\pi}$, $s_{k+1} \sim P(\cdot|s_k,\pi(s_k))$, and α_k is the learning rate (or step-size). Note that the update is done directly after witnessing the transition $(s_k, a_k = \pi(s_k), r_{k+1}, s_{k+1})$.

The notation d^{π} above denotes the stationary state distribution under policy π. The update term, $R(s_k,\pi(s_k)) + \gamma V_k(s_{k+1}) - V_k(s_k)$, is called the TD error, and it measures the difference between the current estimated value $V_k(s_k)$ and the improved estimate $R(s_k,\pi(s_k)) + \gamma V_k(s_{k+1})$. For any fixed policy π, TD update converges to V^{π} almost surely (i.e., with probability 1) if the step-size satisfies the so-called *Robbins-Monro rule*, $\sum_{k=0}^{\infty}\alpha_k = \infty$, $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$.

SARSA "SARSA" refers to the procedure of updating Q-value by following a sequence of experience $\ldots, s_k, a_k, r_{k+1}, s_{k+1}, a_{k+1}, \ldots$, and the update rule is given by (10):

$$Q_{k+1}(s_k,a_k) = Q_k(s_k,a_k) + \alpha_k(R(s_k,a_k) + \gamma Q_k(s_{k+1},\pi(s_{k+1})) - Q_k(s_k,a_k)). \tag{10}$$

SARSA runs TD-learning to evaluate the state-action value function Q^{π} corresponding to the current policy π, computes an improved policy using Q^{π}, and alternates both steps to find Q^*. SARSA is an on-policy method because the actions a_k and a_{k+1} used in the update equation are both derived from the policy that is being followed at the time of the update.

Q-Learning The Q-learning algorithms obey the following update rule (11):

$$Q_{k+1}(s_k,a_k) = Q_k(s_k,a_k) + \alpha_k\left(R(s_k,a_k) + \gamma\max_{a\in\mathscr{A}} Q_k(s_{k+1},a) - Q_k(s_k,a_k)\right), \tag{11}$$

where $s_k \sim d^{\pi}$, $s_{k+1} \sim P(\cdot|s_k,\pi^b(s_k))$, and π^b is called the behavior policy, which usually refers to the policy used to collect observations for learning. The algorithm converges to Q^* almost surely [39] provided

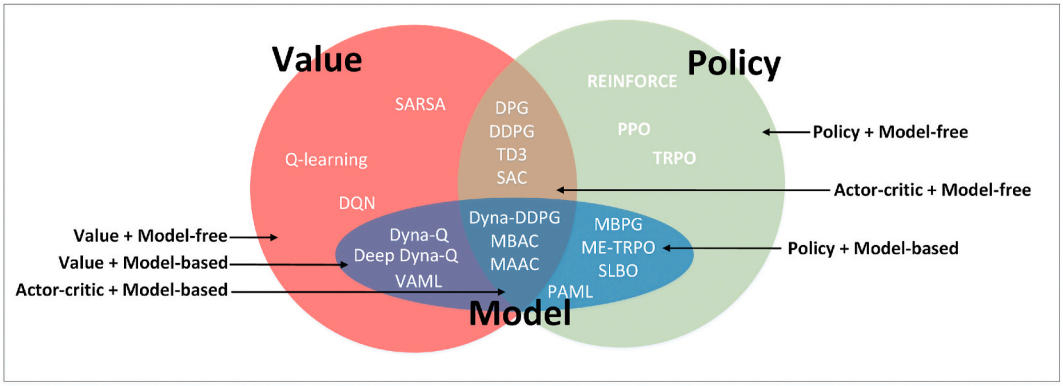

Fig. 6. Classification of RL algorithms (inspired from: UCL Course on RL by David Silver [41]).

that the step-size satisfies the Robbins-Monro rule and every state is visited infinitely often.

SARSA and *Q*-learning are distinct in the way how they evaluate the policy they are optimizing. SARSA evaluates the policy based on the experience from the policy itself whereas *Q*-learning evaluates the policy based on the experience from any behavior policy. Hence, we can use a database of past experiences in *Q*-learning, also referred to as the experience replay buffer. In stark contrast, we have to create a new experience each time a policy is updated for SARSA. As a result, SARSA is called as an on-policy method whereas *Q*-learning is referred to as an off-policy method.

Deep Q-Network (DQN) The function approximators are used to scale up the above lookup table methods to problems with very large state and/or action spaces. The deep *Q*-network (DQN) algorithm [33] is a variant of Q-learning based on neural network (NN) approximations of the Q-function. In particular, we train an NN as parameterized by θ to minimize the following loss function $L(\theta)$, (12):

$$L(\theta) := \mathbb{E}_{(s,a,r,s')\sim\mathscr{D}}\left[\left(\underbrace{r(s,a)+\gamma\max_{a'\in\mathscr{A}}Q(s',a';\theta^{-})}_{\text{Target}}-\underbrace{Q(s,a;\theta)}_{\text{Prediction}}\right)^{2}\right], \tag{12}$$

where $\mathscr{D} := (e_0, e_1, \ldots, e_{l-1})$ is an experience replay buffer of some user chosen length l, which stores the agent's past experience $e_i := (s_i, a_i, r_i, s_{i+1})$ to reduce correlations between observations; and θ, θ^{-} are called the online and target variables, respectively. Stochastic gradient descent steps are taken while freezing the target variable θ^{-}, and the target variable is replaced with the online variable θ periodically after a number of stochastic gradient steps. The experience replay and target network significantly improve and stabilize the training procedure of Q-learning.

There are many extensions of DQN to improve the original design, such as Double DQN [52] and Dueling DQN [53]. The max operator in DQN uses the same network values both to select and to evaluate an action. Thus the DQN algorithm suffers from a substantial overestimation of the value function. Double DQN addresses this issue. Like in DQN, the dueling network is also a DNN function approximator for learning the Q-function. Differently, it approximates the Q-function by decoupling the value function and the advantage function.

3.2.2. Policy based methods

Policy-based methods learn the policy directly with a parameterized function respect to θ, $\pi(a|s;\theta)$. Compared to value-based methods, policy-based methods are effective in continuous action spaces, and they can learn stochastic policies. We consider the following objective (13):

$$J(\theta) := \sum_{s\in\mathscr{S}} d^{\pi_\theta}(s)\sum_{a\in\mathscr{A}}\pi_\theta(a|s)Q^{\pi_\theta}(s,a), \tag{13}$$

where $d^{\pi_\theta}(s)$ is stationary distribution of Markov chain for π_θ. The policy gradient theorem, which lays the theoretical foundation for various policy gradient algorithms, states that the gradient of the above objective is given by (14)

$$\nabla J(\theta) = \mathop{\mathbb{E}}_{\pi_\theta}\left[\nabla \ln \pi_\theta(a|s) Q^{\pi_\theta}(s,a)\right] = \mathop{\mathbb{E}}_{\pi_\theta}\left[G_t \nabla \ln \pi_\theta(A_t|S_t)\right]. \tag{14}$$

Here G_t is the discounted cumulative return starting from time step t. Policy gradient methods search for a local maximum in $J(\theta)$ by ascending the gradient of the policy, w.r.t parameters θ. Below we discuss three prominent policy-based methods: REINFORCE, Trust region policy optimization (TRPO), and Proximal Policy Optimization (PPO) (cf. Fig. 6).

REINFORCE REINFORCE, also known as Monte-Carlo policy gradient, relies on $Q^{\pi}(s,a)$, an estimated return by MC methods using episode samples, to update the policy parameter θ (15):

$$\theta \quad \leftarrow \quad \theta + \alpha\gamma^{t} G_t \nabla \ln \pi_\theta(A_t|S_t) \tag{15}$$

The use of stochastic gradient method ensures the convergence to a local optimum when choosing a decreasing step α_t such that $\sum_{k=0}^{\infty}\alpha_k = \infty$, $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$.

The above vanilla policy gradient update has no bias but high variance. A widely used variation of REINFORCE is to subtract a baseline value from the return G_t to reduce the variance of gradient estimation while leaving the expected value of the update unchanged. For example, a common baseline is to subtract state-value from action-value.

Trust region policy optimization (TRPO) To improve training stability, we should avoid parameter updates that change the policy too much at one step. The key idea in TRPO [54] is to define a KL divergence based trust region that constrains updates to the policy. This constraint is in the policy space rather than in the parameter space and becomes the new "step size" of the algorithm. In this way, we can approximately ensure that the new policy after the policy update performs better than the old policy. Concretely, TRPO solves the following optimization problem (16)-(19):

$$\max_{\pi'} L_{\pi}(\pi') \tag{16}$$

$$\text{s.t.} \quad \overline{D}_{\mathrm{KL}}(\pi,\pi') \le \delta, \tag{17}$$

where

$$L_{\pi}(\pi') = \mathop{\mathbb{E}}_{s\sim d^{\pi},a\sim\pi(\cdot|s)}\left[\frac{\pi'(a|s)}{\pi(a|s)}\{Q^{\pi}(s,a)-V^{\pi}(s)\}\right], \tag{18}$$

$$\overline{D}_{\mathrm{KL}}(\pi,\pi') = \mathop{\mathbb{E}}_{s\sim d^{\pi}}[D_{\mathrm{KL}}(\pi,\pi')[s]], \tag{19}$$

and δ is a hyperparameter.

Proximal Policy Optimization (PPO) PPO relies on a clipped surrogate objective function to ensure that the new policy does not get far from the old policy. PPO is significantly simpler to implement, and empirically seems to perform at least as well as TRPO. Denote the probability ratio between old and new policies as (20):

$$r_k(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}. \tag{20}$$

PPO imposes the constraint by forcing $r_k(\theta)$ to stay within a small interval around 1, precisely $[1-\varepsilon, 1+\varepsilon]$, where ε is a hyperparameter. PPO updates policies via (21)

$$\theta_{k+1} = \arg\max_\theta \mathop{\mathbb{E}}_{s,a\sim\pi_{\theta_k}} [L(s,a,\theta_k,\theta)] \tag{21}$$

typically taking multiple steps of (usually minibatch) SGD to maximize the objective. Here L is given by (22)

$$L(s,a,\theta_k,\theta) = \min\{r_k(\theta)A^{\pi_{\theta_k}}(s,a), \text{clip}(r_k(\theta), 1-\varepsilon, 1+\varepsilon)A^{\pi_{\theta_k}}(s,a)\}. \tag{22}$$

The function $\text{clip}(r_k(\theta), 1-\varepsilon, 1+\varepsilon)$ clips the ratio within $[1-\varepsilon, 1+\varepsilon]$.

3.2.3. Actor-critic methods

Policy-based (actor only) methods explicitly learn a policy that implicitly maximizes the discounted cumulative reward; however, these methods are disadvantaged by high variance in gradient estimates. Value-based (critic only) methods use the Bellman optimality relationship to derive policy from the learned value function; these methods have a lower variance in expected return estimates. The actor-critic methods combine the advantages of the actor-only and critic-only methods. The actor-critic algorithms parameterize both policy and value functions and simultaneously update these in training; thus they belong to both value-based and policy-based groups (cf. Fig. 6). They often exhibit better empirical performances than the value-based only and policy-based only methods.

Deterministic policy gradient (DPG) In an off-policy setting, we consider the following performance objective (23):

$$J_b(\mu) = \int_{\mathscr{S}} d^b(s) Q^\mu(s, \mu(s)) ds, \tag{23}$$

which is the value function of the target policy μ, averaged over the state distribution of a (stochastic) behavior policy $b : \mathscr{S} \times \mathscr{A} \to [0,1]$. Then, we consider parameterized deterministic policies $\{\mu_\theta : \theta \in \Theta\}$, and search for θ to maximize the performance objective $J_b(\mu_\theta)$. By an abuse of notation, we denote $J_b(\theta) = J_b(\mu_\theta)$. If the policy parametrization μ_θ is differentiable, under some regularity conditions on the MDP, the gradient of $J_b(\theta)$ can be expressed as (24)

$$\nabla_\theta J_b(\theta) \approx \mathop{\mathbb{E}}_{s\sim\rho^b} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}\right]. \tag{24}$$

The above equation is referred to as *off-policy deterministic policy gradient* [55].

Deep Deterministic policy gradient (DDPG): DDPG [35] is an off-policy method, and it is applicable to continuous action spaces. DDPG is an actor-critic method that maintains a parameterized actor function μ_θ which specifies the current deterministic policy. The critic $Q^w(s,a)$ is parameterized by w and learned using the Bellman equation, by minimizing the empirical Bellman residual $L(w) = \mathop{\mathbb{E}}_{s_t\sim\rho^b, a_t\sim b}[(Q^w(s_t,a_t) - y_t)^2]$, where $y_t = r_t + \gamma Q^w(s_{t+1}, \mu_\theta(s_{t+1}))$, and b is a behavior policy used to collect samples. The workflow of DDPG algorithm is shown in Fig. 7.

Twin Delayed DDPG (TD3): It is well known that Q-learning and deterministic policy gradients suffer from an overestimation bias due to the noise in the value estimates. To address this issue [56], proposed a variant of DDPG built on Double Q-learning, by making the following changes:

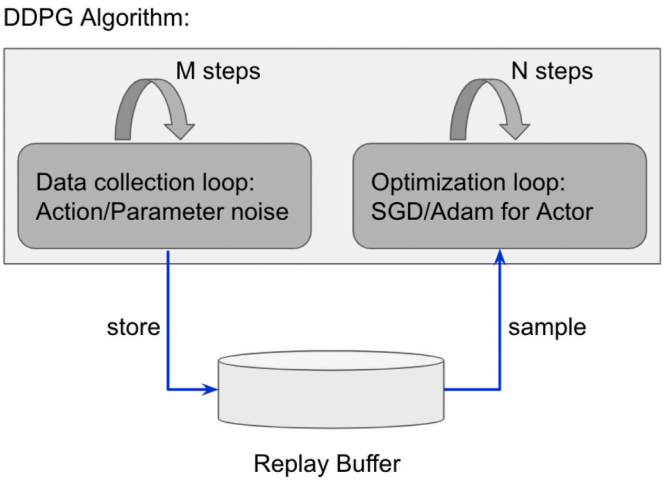

Fig. 7. DDPG algorithm workflow.

1. Maintains a pair of critics along with a single actor.
2. Clipped action exploration: noise added like DDPG but bounded to fixed range as follows (25)

$$a'(s') = \text{clip}\left(\mu_{\theta_{\text{targ}}}(s') + \text{clip}(\varepsilon, -c, c), a_{\text{Low}}, a_{\text{High}}\right), \quad \varepsilon \sim \mathscr{N}(0, \bar{\sigma}) \tag{25}$$

3. Updates of Critic are more frequent than of policy.

Soft Actor Critic (SAC) SAC [57] algorithm incorporates the entropy measure of the policy into the reward to encourage exploration. It is an off-policy actor-critic model following the maximum entropy RL framework. SAC algorithm searches for policy (26):

$$\pi^* = \arg\max_\pi \mathop{\mathbb{E}}_{\tau\sim\pi} \left[\sum_{t=0}^{\infty} \gamma^t \{R(s_{t+1}) + \alpha \mathscr{H}(\pi(\cdot|s_t))\}\right], \tag{26}$$

where $\mathscr{H}\left(\pi(\cdot|s)) = \mathop{\mathbb{E}}_{a\sim\pi(s)}[-\log \pi(a|s)\right]$, and α is the trade-off between reward and entropy. The entropy maximization leads to policies that can (1) explore more and (2) capture multiple modes of near-optimal strategies.

For a detailed review on policy gradient and actor-critic methods, confer [58].

3.2.4. Model-based RL

Although model-free approaches are successfully applied in several domains, such as games and robotics, in practice, their high sample complexity is a critical problem. Model-based approaches have long been recognized as a potential avenue for reducing the sample complexity of RL algorithms. In the model-based RL, the agent interacts with the environment and gathers experience to learn its model. Recently, in Ref. [59], it is demonstrated that model-based methods can be more sample-efficient exponentially than model-free methods in general contextual decision processes. In linear quadratic regulators, a gap between model-based and specific model-free algorithm is reported [60].

The Dyna algorithm [42] alternates between learning the model based on the gathered data that executes the current policy on the environment and improving the policy with imaginary data from the learned model. The Model-Ensemble Trust-Region Policy Optimization (ME-TRPO) [61] is a Dyna-style algorithm, which maintains an ensemble of neural networks to model the dynamics. It also uses the TRPO in the policy improvement step with the data generated by learned dynamics models. Recently, Luo et al. [62] proposed an algorithmic framework for learning both policy and model with a monotonic improvement guarantee; their proposed final practical algorithm (i.e.,

Stochastic Lower Bound Optimization (SLBO)) is a variant of ME-TRPO. Apart from Dyna-style algorithms, there are other types of model-based RL algorithms (cf [63]).

3.2.5. Batch RL

RL algorithms discussed above are all interaction-based methods (online RL), i.e., they interact with the environment to collect data while updating the policy or value parameters. The batch RL algorithms decouple data collection and policy optimization. This means that they operate on a fixed set of experience $\{(s, a, r, s')\}$ collected using certain behavioral policies; however, they do not interact with the environment during policy training [64]. Batch algorithms therefore ignore the exploration-exploitation problem and leverage the data, that is, they are more data-efficient.

The least-squares policy iteration (LSPI) [65] is a model-free batch RL algorithm that alternates between policy evaluation (learning a linear Q-function approximation) and policy improvement steps. The fitted Q-iteration (FQI) [66,67] is the most popular algorithm in batch RL and is a considerably straightforward batch version of Q-learning that allows the use of any function approximator for the Q-function (e.g., random forests and deep neural networks). Here, some recent batch deep RL algorithms, such as random ensemble mixture (REM) [68], batch-constrained deep Q-learning (BCQ) [69], and bootstrapping error accumulation reduction Q-learning (BEAR-QL) [70], are also mentioned.

3.3. Implementation

Several open-source deep RL packages are available with either Tensorflow or PyTorch implementation of state of the art deep RL algorithms and good documentation. The prominent ones include OpenAI Baselines [71], Stable Baselines [72], Tensorforce [73], and MushroomRL [74]. These libraries enable quick and reliable implementation and testing of RL models for the problem of interest. [75] prescribes a list of best practices for the deep RL researchers and practitioners to follow when working on the deep RL experiments and reporting the results.

4. Top-down view of applications

Several RL methodologies have been successfully applied in the energy system domain. Regardless of the methodology or application, two driving factors can be considered as the main reasons for selecting RL over other available methods: RL 1) is effective in handling uncertainties and 2) is a model-free approach. For example, the uncertainties in demand, such as renewable energy potential, energy market, and grid conditions are regarded as major problems in energy dispatch. These uncertainties can be effectively managed by using RL [32], thus making it an effective technique. In other cases, it is difficult to develop a comprehensive model for energy flow; the model-free nature of RL is advantageous in these contexts, such as BEMSs. As explained in Section 3, different RL approaches have been used to overcome these problems. In this section, the effectiveness of these approaches in different domains is assessed in a more holistic manner using a top-down approach. In view of this, Section 4.1 discusses the diversity of the foregoing problems. Section 4.2 provides a cross comparison among the different classes of the problems highlighting the similarities and the difference. Based on that, Section 4.3 elaborates on the different RL methodologies used in the energy system domain. Finally, Section 4.4 explicates the verification methodology, reproducibility of existing state of the art applications.

4.1. Problem diversity

As explained in Section 3, RL can be effectively used in a number of different domains, some of which can be facilely interlinked. This section provides a more holistic overview about the diversity of the control problem addressed in each domain.

4.1.1. Building energy management systems (BEMS)

Building energy management systems (BEMS) mainly focused at maintaining controlling the energy flows within the building while taking into account the changes that take place in climate, occupancy, equipment usage, etc. RL has been used to control heating, ventilation, air conditioning (HVAC), lighting, and blinds considering several objectives, such as minimizing the operational cost, and improving energy efficiency, comfort, and indoor air quality. A number of the publications related to BEMS, including the different elements of the problem, are summarized in Table 1. It is observed that the majority of publications (more than 80%) focus on the optimal use of HVAC systems (Table 1), whereas the number of papers that mainly discuss building elements, such as blinds and lighting appliances, are limited. Considering energy consumption and visual and thermal comforts, the lighting control and HVAC of a building can be effectively linked. Except for the papers of Park et al. [76] and Cheng et al. [77], the above aspects have not been reported; only lighting has been discussed. Similarly, although the interactions with the grid have an important role in the HVAC system operation, these are only presented in four papers. The indoor air quality has been elaborated only by Ref. [78–80]. As for minimizing operational cost, only Wen et al. reported on the scheduling of building jobs [81]. It can thus be concluded that the majority of RL applications are only focused on improving the thermal performances of BEMS. Accordingly, by considering several building elements, the potential of improving the application diversity is considerable. Although cost minimization is a major problem, most of the publications in BEMS only discuss energy efficiency and visual or thermal comfort. Numerous studies focus on several objectives where weighted objective functions are considered. It can be concluded that a specific control problem is considered in the BEMS sector where there is ample opportunity to improve the diversity.

4.1.2. Dispatch problem

The dispatch problem is one that is widely examined in the energy system domain. Irrespective of the methodology, the dispatch problem is well investigated because of the popularity of renewable energy technologies and energy storage. Several notable changes can be observed when BEMS is shifted to the dispatch problem. According to the list in Table 2, focus shifts from heat to electricity, and cost is given a higher priority in the dispatch problem. Except for [82–84], none of the papers gave a report on heating. Kofinas et al. [85,86] considered desalination as an application of the dispatch problem, but all other papers (95%) only focused on the electricity sector. Furthermore, more than 80% of articles focused on minimizing system cost, again indicating a notable deviation from BEMS. A considerably broader problem with a number of control elements is considered in the dispatch problem. The optimal operation of energy storage and dispatchable sources are considered by 82% and 60% of the publications, respectively. The dispatch strategy is sensitive to fluctuations in renewable energy sources, as reported in more than 70% of the papers. Similarly, the factors that influence grid price signals are considered by more than 50% of the publications. Evidently, compared to BEMS, a more complex problem is taken into account in the dispatch problem, and more diversified system designs are considered according to their application. A moderate number of publications use multi-agent reinforcement learning (MARL). Although the majority of papers report that a cooperative scenario can lead to a correlated equilibrium, its problem formulation is usually more exigent than that of a single-agent scenario. Foruzan et al. [87] considered the non-cooperative scenario where the Nash equilibrium is guaranteed, making it a more extensive process than the single-agent RL. Finally, it can be concluded that RL is well established as a potential method for solving the dispatch problem. More diverse operation problems have been solved considering numerous elements in an energy system. It will be interesting to consider heating, cooling, and other energy services as well as the dispatch problem for future research.

Table 1
Elements of the BEMS considered in the present state of the art.

	HVAC	Building Elements	Energy Storage	Lighting	Indoor Air Quality	Job Scheduling	Price Signals	Objective: Comfort	Objective: Cost	Objective: Energy Efficiency
Eller et al. [78]	✓				✓			✓		✓
Chen et al. [96]	✓	✓						✓		✓
Peirelinck et al. [97]	✓								✓	
Kazmi et al. [98]			✓							✓
Valladares et al. [79]	✓				✓			✓		
Zhang et al. [99]	✓									✓
Heo et al. [80]	✓				✓					✓
Liu et al. [100]	✓									✓
Yoon and Moon [101]	✓							✓		✓
wei et al. [102]	✓						✓		✓	
Yu and Dextor [103]	✓						✓	✓	✓	
Wen et al. [81]						✓			✓	
Gracia et al. [104]		✓	✓							✓
Wang et al. [105]	✓							✓		✓
Ruelens et al. [106]			✓				✓		✓	
Ruelens et al. [107]	✓							✓		
Jia et al. [108]	✓							✓		✓
Kazmi et al. [109]	✓		✓				✓			✓
Park et al. [76]	✓			✓				✓		✓
Cheng et al. [77]	✓	✓		✓				✓		
Schmidt et al. [110]	✓							✓		✓
Kazmi et al. [111]	✓		✓							✓

4.1.3. Energy markets and grid

The number of RL applications beyond BEMS and dispatch problems is steadily increasing. The participation in energy markets–day ahead market, balancing market, or spot market–has been performed by using RL. All papers related to energy markets only focus on electricity (Table 3). The optimal control strategies for energy storage, dispatchable source, and demand response including fluctuations in the energy markets are obtained by using RL. Compared to the dispatch problem, a simplified energy system, which can be extended to consider more comprehensive problems, is considered in these studies. However, 44% of the papers consider MARL, which is a reasonable improvement when shifting from both BEMS and dispatch problems, allowing the participation of multiple sectors. This indicates that a theoretical platform already exists for a broader extension of energy market problems by being linked with other problems, such as dispatch and BEMS.

A grid facilitates the linking of energy systems with other energy systems and markets. Therefore, it performs an important role when considering RL applications of energy systems; it can also be considered as a boundary. RL has been appropriately used for grid operation from low to high voltages taking into account several aspects (Table 4). This study does not attempt to review the publications that report the use of RL to resolve the grid operation problem; instead, a few papers that discuss the interaction between energy systems and the grid are selected. The publications that present the link between energy systems and grid consider two aspects: reliable operation and stability of grid linked with generation. The papers that discuss reliable operation consider the n-1 security (Zarrabian et al.) [88] and component outage caused by maintenance activities (Rocchetta et al.) [89]. In some papers, the stability aspect is considered, including the frequency/voltage stability and transient stability of the grid. However, it remains problematic to include the optimal operation of the energy system and the healthy operation of the grid; these will be interesting research directions for future investigations.

4.1.4. Vehicles and energy devices

The applications of hybrid vehicles and energy devices have a relatively specific scope compared to that presented in Section 4.1.1–3. RL algorithms that are used in the energy management systems of vehicles either have multiple storage devices or an energy storage device with an internal combustion engine (ICE). The energy management problem of vehicles is similar to the dispatch problem discussed in Section 4.1.2. The operational strategy varies depending on the traffic and driving style of vehicles. Most publications focus on the combination of ICE and battery bank (Table 5), and a few of them report on the combination of a battery bank with H_2 and supercapacitor storage. Furthermore, the majority of publications focus on improving fuel efficiency; for example, Ermon et al. [90], Xiong et al. [91], and Reddy et al. [92] successfully improved the battery lifetime. Brusey et al. [93] used RL to improve the thermal comfort inside vehicles and formulated a problem similar to the BEMS. However, the link between transportation and electric charging that is usually discussed with the BEMS or dispatch problems has not been considered in any of these publications, and the details on energy flow within a vehicle are not considered in depth as in other publications. It is further observed that most of the papers are published by only one research group.

Similar to energy systems in vehicles, RL applications in energy devices are mainly related to the optimal power point tracking for renewable energy devices. The maximum power point tracking (MPPT) for wind turbines, solar panels, and wave energy generators are considered in this context (Table 6). The main difference between MPPT, typical BEMS, and dispatch problem is that the MPPT should be considered at a finer time resolution than other problems. Usually, in the MPPT problem, the operation should be in the scale of seconds, whereas the BEMS and dispatch problem are solved at a resolution of 15 min or 1

Table 2
Diversity of the elements considered for the dispatch problem. Much broader problems are considered under the dispatch problem mainly focusing on the electricity sector.

	Electricity	Heating	Elec. Storage	Heat Storage	Desalination	Dispatchable	Renewables	Price Signals	Multi Agent RL	Demand Response	Cost	Other Rewards	Emissions
Hua et al. [112]	✓		✓			✓	✓	✓					
Zhou et al. [113]	✓					✓	✓				✓		✓
Bollenbacher & Rhein [82]	✓	✓	✓			✓					✓		
Kofinas et al. [85]	✓		✓		✓		✓					✓	
Sun et al. [83]	✓	✓	✓			✓	✓				✓		✓
Ji et al. [114]	✓		✓			✓	✓	✓			✓		
Menon et al. [115]	✓		✓			✓	✓		✓		✓		
Sheikhi et al. [116]			✓	✓		✓		✓			✓		
Yu et al. [117]	✓					✓			✓		✓		
Berlink et al. [118]	✓		✓					✓		✓	✓		
Wang et al. [119]	✓		✓				✓	✓		✓	✓		
Yu et al. [120]	✓					✓	✓		✓	✓	✓	✓	
Venayagamoorthy et al. [121]	✓		✓			✓	✓	✓					
Guan et al. [122]	✓		✓				✓	✓			✓		
Abdulla et al. [123]	✓		✓					✓			✓		
Foruzan et al. [87]	✓		✓			✓	✓	✓	✓	✓	✓		
Liu et al. [124]	✓		✓			✓	✓	✓	✓		✓		
Bukenberger & Palmintier [125]	✓					✓	✓				✓		
Ebell & Pruckner [126]	✓		✓				✓	✓	✓		✓		
Qiu et al. [127]	✓		✓				✓				✓		
Shirzeh et al. [128]	✓		✓				✓				✓		
Kofinas et al. [86]	✓		✓		✓		✓		✓	✓		✓	
Mbuwir et al. [84]	✓	✓	✓			✓	✓	✓			✓		

h. As a result, no study that uses RL to link energy system operation and MPPT is found although these two problems are closely related (see Fig. 8).

4.2. Cross comparison among different topics and studies covering several areas

The operation problems discussed under Section 4.1 are having many similarities. As shown in Fig. 10, dividing each problem class into sub-areas makes it easy to understand the interrelations. For example, BEMS is closely linked with the dispatch problem, where many areas have been shared commonly. For example, demand response, integration of renewables, energy storage, energy demand for heating and cooling, and overall cost reduction in the operation have been commonly discussed in these two classes. This is quite clear when analyzing the integrated class of problems, as shown in Fig. 9. Out of the total publications, 70% of the publications consider BEMS the dispatch problem in the integrated class. Most of the publications focus on cost minimization while considering the fluctuations in the grid electricity price, renewable energy generation, and demand (Table 7). Similarly, electric vehicle charging has been closely linked with both BEMS and dispatch problems. Such a detailed workflow that covers several sectors shows the potential of RL to link several energy management problems; this is a relatively exciting attribute of the energy transition where sector coupling is expected to perform a major function. However, reasonable simplification concerning the BEMS and dispatch problem is observed when extending the scope of the problem. For example, most of these publications do not pay careful attention to the building energy model, energy conversion process in the system components, etc. Furthermore, BEMS is having relatively low interactions with the energy market and grid, which needs to be further improved interlink the building sector to the energy internet.

The dispatching problem is having a close relationship between BEMS, Market, and Grid besides maintaining close interlinks between the BEMS. The close relationship between dispatch and market is easy to understand as energy systems' operation is often controlled by the price signals from the energy markets. Several studies use RL for storage management, renewable energy integration taking into account the energy markets that are broadly discussed in Section 4.1.2. Often these studies consider price signals from the energy markets. However, some studies consider the detailed dynamics of the energy markets and dispatch, as presented in Section 4.1.3. Simultaneously, multi-agent models have been introduced to represent the participation of different micro-grids and different components of the same micro-grid in energy markets. Claessens et al. [94] and Foruzan et al. [87] used MARL in developing an energy management strategy, which is a more difficult task. Linking several sectors can present more complexities related to uncertainties and model approximation. Similar to BEMS, the dispatch problem is oversimplified when moving into the multi-agent systems when dealing with the energy markets. It is not easy to find publications that cover a reasonable level of physics (depth) while accommodating the dispatch and energy markets. Energy markets are going through significant expansion, opening into many participants. Energy markets with large-scale participation where a large group of distributed energy systems interacts are not discussed (where energy markets need to be considered and the grid and dispatch problem together).

The relationship between the grid and the energy system can be understood as the dispatch strategy facilitates to maintain grid stability, which leads to a coupling between the optimal power flow and optimal dispatch problems. This is clear when analyzing the publications in the dispatch domain (4.1.2). As it was bound in the building sector, a reasonable simplification of the control problem is observed when linking these two problems. Machine learning techniques such as supervised learning (especially with graph convolution) have been used for optimal power flow problems more often than the use of reinforcement learning. Therefore, joint consideration of the optimal power-flow

Table 3
Elements considered for RL problems in energy markets.

	Electricity	Renewables	Storage	Dispatchable	Day Ahead Market	Spot Market	Demand Response	Optimal Power Flow	Multi Agent	Nash Eqi.
Naghibi-Sistani et al. [129]	✓			✓		✓			✓	✓
Nanduri & Kazemzadeh [130]	✓			✓		✓		✓	✓	
Chen & Su [131]	✓	✓	✓			✓				
Lu et al. [132]	✓				✓		✓			
Lu & Hong [133]	✓					✓	✓			
Kian et al. [134]	✓			✓	✓			✓		
Igushi et al. [135]	✓		✓		✓				✓	
Kim et al. [136]	✓					✓	✓		✓	
Peters et al. [137]	✓				✓	✓	✓			

Table 4
Elements considered for RL problems in Grid.

	health state of the grid	Voltage/frequency regulation	Transient stability	Renewables	Storage	Dispatchable	Security	Cyber connection	Multi agent
Rocchetta et al. [89]	✓								
Carvalho et al. [138]				✓	✓			✓	✓
Mahmoud et al. [139]		✓							
Cao et al. [140]		✓	✓						
Sun et al. [141]		✓		✓	✓			✓	✓
Zarrabian et al. [88]						✓	✓		

Table 5
Applications of RL in Vehicle energy systems.

	Battery	Super cap/ultra cap	H2	Engine	Battery losses	Fuel economy	Objective function
Ermon et al. [90]	✓	✓			✓		Battery losses
Liu et al. [10]	✓			✓		✓	Fuel consumption by engine
Liu et al. [142]	✓			✓		✓	Fuel consumption by engine
Liu et al. [143]	✓			✓		✓	Fuel consumption by engine
Zou et al. [144]	✓			✓		✓	Fuel consumption by engine
Yuan et al. [145]	✓		✓			✓	Hydrogen consumption
Xiong et al. [91]	✓	✓			✓		Battery energy loss, ultra-capacitor energy loss and DC/DC converter loss
Brusey et al. [93]	✓						Thermal comfort
Reddy et al. [92]	✓		✓		✓		Improving battery life by minimizing charge discharge cycles
Wu et al. [146]	✓			✓		✓	Sum of cost for the fuel and electricity consumption
Zhou et al. [147]	✓			✓			Energy used for vehicle operation
Qi et al. [148]	✓			✓		✓	Fuel efficiency

Table 6
Applications of RL in Energy system devices.

	Solar MPPT	With Shading	Wind	Yaw control	Wave energy
Kofinas et al. [149]	✓				
Anderlini et al. [150]					✓
Wie et al. [151]			✓	✓	
Zhang et al. [152]	✓	✓			
Aguirre et al. [153]			✓	✓	

problem and the dispatch strategy using RL is not found in the literature. Often, the influence of micro-grids on the local grid stability and the healthy operation of the grid are taken into account along with the dispatch problem. It would be interesting to evaluate the potential of reinforcement learning to consider the dispatch and grid problems jointly that will notably help peer-to-peer trading, frequency regulation markets etc. Network multi agent architecture might be quite interesting in this regard, which has been less discussed within the domain.

A limited number of publications attend to link several sectors. Most of the publications covering a broad set of domains are centered on the dispatch problem that often focuses on the electricity sector. This highlights the dispatch problem's capability to be the hub of while facilitating other domains to be well connected to it. Secondly, BEMS have been closely linked with the dispatch problem within the integrated problem domain. This highlights RL's potential to extend the boundaries of classical dispatch or unit commitment problem that is often limited to the generation. RL enables considering the complex domains of the energy demand such as lighting, user comfort etc. (Lee & Choi [95]) along with the dispatch problem that brings energy system operation to become more user-friendly and adaptive. Similarly, the dispatch problem can be easily linked with the vehicle to grid problem, which is quite similar to the job shop scheduling problem. This

Table 7
Applications of RL in integrated Problems.

	BEMS	Dispatch	Market	Transportation	Electricity	Heating	Cooling	EV	Lighting	Elec. Sto.	Thermal Sto.	Dispatchable sources	Renewables	Price signals	Energy market	Multi agent	Demand response	Energy efficiency	Comfort/ Satisfaction	cost
Lee & Choi [95]	✓	✓			✓	✓				✓			✓	✓			✓		✓	✓
Remani et al. [154]	✓	✓			✓	✓			✓	✓			✓	✓			✓			✓
Henze & Schoenmann [155]	✓	✓				✓	✓				✓									✓
Claessens et al. [94]	✓	✓				✓						✓			✓	✓		✓		✓
Canteli et al. [156]	✓	✓			✓		✓				✓		✓	✓						
Foruzan et al. [87]		✓	✓		✓					✓		✓	✓		✓	✓		✓		✓
Kim & Lim	✓	✓		✓	✓			✓		✓			✓	✓			✓			✓
Rayati et al. [157]	✓	✓		✓	✓	✓		✓		✓		✓	✓	✓			✓		✓	✓
Mocanu et al. [158]	✓	✓		✓	✓			✓												✓
Odonkor & Lewis [159]	✓	✓	✓		✓					✓										✓
Tang et al. [160]		✓		✓	✓			✓				✓	✓	✓						✓
Ko et al. [161]		✓		✓	✓			✓											✓	
Dang et al. [162]		✓		✓	✓			✓						✓						✓

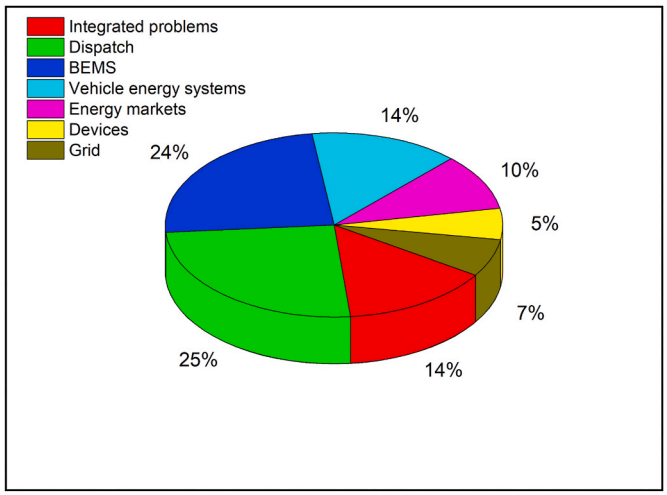

Fig. 8. Pie chart of publications on various sectors; more publications are focused on specific problems (SP) than on integrated problems (IP).

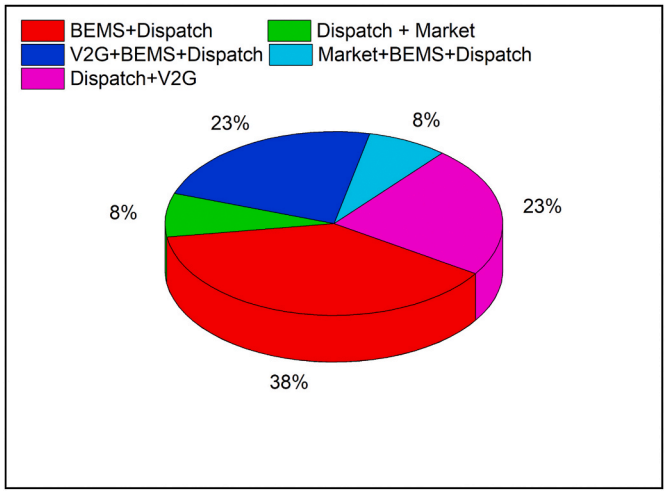

Fig. 9. Classification of integrated problems.

highlights RL's capability to link the manufacturing sector into the energy markets while considering complex operations within the industrial sector, an interesting future extension. However, the interlinks between vehicle and device domains are limited. Only a limited number of links between the grid operation and MPPT of devices are found with other fundamental problems. It can be considered that this is mainly because of the mismatch among the time resolutions. Grid operation and MPPT both demand a considerably finer time resolution than the other problems. It would be interesting to explore potential means to link these sectors while resolving the mismatch problems among time resolutions.

4.3. RL methodologies

RL methods can be classified in several different ways. In this study, RL techniques are grouped into three main classes: interaction-based methods, interaction-free methods (termed as batch RL), and other methods, such as gradient-free techniques. Based on the state of the art, the majority of publications (more than 80%) report the use interaction-based methods. A detailed evaluation of the applicability of these techniques is presented in Section 4.2.1. Interaction-free methods include batch learning, which appears promising for energy systems because of the availability of historical data, as discussed in Section 4.2.2. Finally, other alternatives, which do not fall under any of the

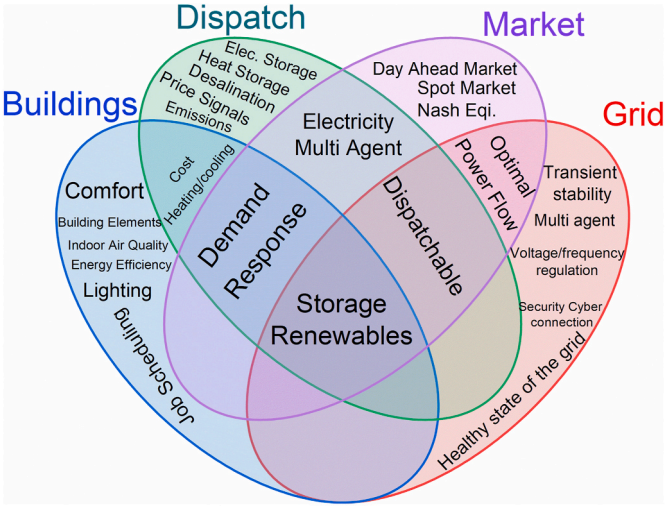

Fig. 10. A Venn diagram presenting areas that are commonly discussed under each problem class. There are many common topics among these problem classes which highlights the importance of a generic approach.

aforementioned categories, are discussed in Section 4.2.3.

4.3.1. Interaction-based methods

The class of interaction-based RL methods include many techniques. RL algorithms that belong to this class interact with the environment during the learning process. The literature on interaction-based methods is categorized as value-based RL, policy-based RL, actor-critic, and model-based RL methods; the foregoing is aligned with the classification presented in Section 3.

Value-based RL. Value-based methods, including Monte Carlo (MC), Q-learning, and SARSA methods, have been extensively used in the energy system domain. Among these three techniques, the Monte Carlo (MC) approach is less preferred by the energy system community because of its sample complexity resulting from the high variance of complete trajectories (Table 8). The MC has been used in only one study (Rayati et al.) [157], i.e., under the IP category in the tabular setting. The SARSA is also a less frequently used approach, and approximately 7% of papers in the SP category report the use of this technique. The authors found no study that uses SARSA to solve integrated problems. Both standard and lambda SARSA methods have been applied in the domains of BEMS [104], dispatch [122], energy markets [137], and vehicle energy systems [148]. A clear contrast is observed in shifting from both MC and SARSA to Q-learning (Table 8).

In the energy system domain, Q-learning is the predominantly employed approach, i.e., more than 50% in the specific problem category and 57% in the integrated problem domain (cf. Table 8). Q-learning has been applied in all energy system domains. For its branches, Q-learning in the tabular setting is used in all domains–BEMS [78,96], dispatch [83,85], energy markets [129,132], grid [88,138], vehicle energy systems [10,142], and devices [149,151]. On the other hand, the use of Q (lambda) algorithm has been reported in publications related to BMES ([103]) and dispatch ([117,120]); it has potentially application in other domains. In addition to tabular Q-learning, different function approximators have been explored, i.e., linear function approximation [124], multi-layer perceptron [115], convolutional neural network [140], and deep neural network (DNN) [79,89,131]. Q-learning with function approximators is typically preferred for the specific problem category. However, note that there is a difference between Q-Learning with DNN and deep Q-network (DQN): in addition to function approximation, the DQN introduces further stratagems, such as experience replay and target network, to improve stability. Only publications related to BMES (SP) and dispatch (SP) have reported the use of the DQN. The use of sophisticated DQN variants, such as double DQN (except [79]), dueling DQN, and rainbow have not been reported in any of the published papers, and some papers have reported the use of different variants of Q-learning, such as the greedy GQ-Algorithm [124], speedy QL algorithm [143], and multistep RL [147]. The following is related to SARSA. In addition to tabular methods, function approximation methods, such as linear function approximation [137] and tile coding [93], have also been considered. It is interesting to note that Ebell and Pruckner [126] considered the MDP setting with both discrete and continuous state spaces.

Policy-based RL. In the literature on energy system (based on RL), the use policy-based methods are less frequently reported (Table 8). In particular, in all the papers considered, only 3.2% in the specific problem group and 3.8% in the integrated problem category have applied policy-based RL methods. Even these few works only utilized the policy-based methods, such as vanilla policy gradient (with and without baseline) [105,108,158] and TRPO [82]. In the selected pool of papers, the authors have not found any work that uses the PPO, which is another state-of-the-art policy-based approach. It is thus concluded that a detailed investigation of the effectiveness of policy-based methods for energy-related problems should be performed.

Actor-critic methods. In the model-free setting, after Q-learning, the actor-critic approach is more frequently applied in energy systems.

Table 8
RL methods for Energy systems (Tab. = Tabular; FA = Function Approximation).

Problem	Q-Learning		SARSA		Monte-Carlo		Policy Gradient	Actor Critic	MBRL + Planning		Batch RL		Evol. RL + Other Methods
	Tab.	FA	Tab.	FA	Tab.	FA			Tab.	FA	Tab.	FA	
BEMS	22.7	18.2	4.5	0.0	0.0	0.0	9.1	9.1	4.5	9.1	4.5	13.6	4.5
Dispatch	39.1	13.0	6.5	2.2	0.0	0.0	2.2	6.5	8.7	4.3	4.3	0.0	13.0
Energy Markets	44.4	11.1	5.6	5.6	0.0	0.0	0.0	11.1	0.0	0.0	0.0	0.0	22.2
Grid	33.3	33.3	0.0	0.0	0.0	0.0	0.0	16.7	16.7	0.0	0.0	0.0	0.0
Vehicle energy systems	50.0	0.0	8.3	8.3	0.0	0.0	0.0	8.3	16.7	0.0	0.0	0.0	8.3
Devices	40.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	0.0	20.0
SP Total	36.4	14.3	5.2	2.6	0.0	0.0	3.2	8.4	7.8	5.2	2.6	3.9	10.4
BEMS + Dispatch	60.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	20.0	0.0
Dispatch + Market	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
V2G + BEMS + Dispatch	33.3	16.7	0.0	0.0	33.3	0.0	16.7	0.0	0.0	0.0	0.0	0.0	0.0
Market + BEMS + Dispatch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
Dispatch + V2G	66.7	0.0	0.0	0.0	0.0	0.0	0.0	33.3	0.0	0.0	0.0	0.0	0.0
IP Total	53.8	3.8	0.0	0.0	7.7	0.0	3.8	15.4	0.0	0.0	7.7	7.7	0.0
Total	38.9	12.8	4.4	2.2	1.1	0.0	3.3	9.4	6.7	4.4	3.3	4.4	8.9

Specifically, 8.4% in the specific problem domain and 15.4% in the integrated problem category have applied actor-critic methods (cf. Table 8). The modern deep actor-critic algorithms, such as A3C [99,112] and DDPG [82,100,146,159], have only been used by the works focusing on BMES (SP), dispatch (SP), vehicle energy systems (SP), and IP. Furthermore, the authors have not encountered any report on work that utilizes state-of-the-art actor-critic algorithms, such as TD3 and SAC, in the pool of papers considered. Finally, it should be noted that despite the successful results achieved by using model-free RL methods, these techniques often require numerous samples.

Model-based RL (MBRL). Interestingly, considerable effort (13%) has been devoted to the specific problem category in applying the model-based RL/planning methods in energy systems, whereas the integrated problem domain has entirely ignored this approach. Compared to specific problems, integrated problems involve complex energy models. Thus, learning an accurate model (with a small error) in the IP setting would require a significant number of samples compared to the SP case. Learning an accurate model is extremely crucial for the MBRL; otherwise, a compounding error problem would result. In this situation, learning local models would be advantageous to the improvement of sample complexity and keeping the modeling error as low as possible [163]. In addition to the tabular representation of transition dynamics, function approximators are also used to represent the model [98,111].

Note that only the studies related to BMES [76,98] and vehicle energy systems [90] have attempted to learn the model and apply RL/planning algorithms to the learned model. In these works, while learning the model, the transition dynamics is represented in tabular form [10,76,90,144], RF [98], NN [98], or DNN ensemble [111]. The Dyna-Q algorithm, which alternates between model-learning and Q-learning blocks, has also been utilized in energy systems [10,144]. It should be noted that a recent deep RL-related paper reports the extensive investigation of MBRL because of this technique's sample-efficient nature [59,60]. The MBRL methods are predominantly employed in the robotics domain [36,164]. The energy system community should also consider these recent advances and adapt them to energy-related problems.

4.3.2. Batch RL

These methods depend on historical data to learn control strategies. Moreover, they do not interact with the environment while learning. The use of batch RL algorithms have only been observed in works related to BEMS (SP), dispatch (SP), and BEMS + dispatch (IP). This method has a significant potential to be used with model predictive control, which can provide a rich historical dataset. In Ref. [84,94,110], batch RL methods have been used in the tabular setting. The fitted Q-iteration algorithm with function approximators, such as DNN [97,156], ERT [106], and ERT ensemble [107] has also been employed.

4.3.3. Other techniques

RL applications that are not categorized under the aforementioned classes are considered under other techniques. Most of the publications report the use non-gradient techniques for optimization and mainly rely on evolutionary algorithms. Evolutionary algorithms have been directly used to train both BEMS and dispatch problems on the rules of Markov decision processes. Both fuzzy [31,165] and crisp rules [166,167] have been considered in this regard. Crisp rules are frequently employed for simple energy systems (e.g., standalone hybrid energy systems), whereas fuzzy rules are used for the dispatch problem of grid-integrated energy systems, which present a considerably complex state space. Although there is a recent trend towards neuro-evolutionary RL, the authors have not encountered any publication that reports the use of evolutionary algorithms to optimize the function approximation process. In addition, hybrid approaches that combine gradient methods and evolutionary algorithms have also not been found. Most of the studies that use evolutionary algorithms consider distributed rewards that are capable of implementing gradient-based methods. It would be interesting to analyze whether the use of evolutionary algorithms in some of those applications affords added advantages. In addition to evolutionary algorithms, the fussy logic and TRPO have been directly applied.

4.3.4. Final remarks

Both BEMS and dispatch-related studies have employed all RL methods, except for the MC method. On the other hand, the BMES + dispatch integrated problem-related studies have only used Q-learning and batch RL methods. In the investigations related to energy markets and vehicle energy systems, Q-learning, SARSA, actor-critic, and evolutionary/other methods, have been employed. Moreover, the studies related to vehicle energy systems used the MBRL methods. In grid-related research, only Q-learning, actor-critic, and MBRL methods are used. In device-related studies, only Q-learning, MBRL, and evolutionary/other methods have been utilized. On the integrated problem side, Foruzan et al. [87] (related to dispatch + market) employed Q-learning; Odonkor and Lewis [159] (related to market + BMES + dispatch) used the actor-critic method. In the V2G + BMES + dispatch category, Q-learning, MC, and PG methods have been applied. Furthermore, the dispatch + V2G-related studies have used Q-learning and actor-critic methods; Q-learning is the most predominantly used approach (more than 50%) in both SP and IP. With the broad range of RL methods, it would be advantageous to investigate other RL methods across all SPs and IPs. It is necessary to create benchmark datasets for energy-related problems so that different RL algorithms can be fairly compared. Model-based RL (specifically learning local models for integrated problems) and batch RL methods seem to be techniques that are worth exploring because they may lead to interesting open research problems in RL literature.

4.4. Performance improvement, verification, and reproducibility

Considering the state of the art, 65% of the studies discussed in Section 4.1 compare the results obtained using RL with an alternative method. The rule-based technique, alternative RL method, fuzzy logic, heuristic models, etc. Are used in this regard. The majority of publications pertaining to the BEMS sector report the use of an alternative method to validate the results. In contrast, fewer studies related to the dispatch problem employ an alternative methodology to validate the results. However, simple approaches, such as set point strategies or simple rule-based techniques, have been used to benchmark RL algorithms in the BEMS sector, achieving a highly promising 10–20% improvement; such a significant performance is not typically observed in other sectors. For example, the performance improvement in the vehicle and device sectors are usually less than 5%. In some cases, RL have been outperformed by the detailed white box approaches, such as model predictive control. It is thus necessary to conduct a cross comparison between more advanced control strategies apart from only using simple-rule based strategies to benchmark the algorithms. In order to create such an environment, it is important to maintain a public repository of these algorithms and problem datasets as that implemented in other communities, such as computer vision. Such repositories will encourage research groups to benchmark RL applications especially for uncommon problems, such as energy system dispatch. Finally, considerably few publications report the use experimental techniques to validate results. Experimental validation requires creating a prototype and implementing RL algorithm in reality. Such an implementation in addition to computational simulation will considerably aid in presenting the potential of RL to a wider range of communities.

5. Future perspectives

This section is devoted to explaining the bottlenecks of the present state-of-the-art methods and identify directions that are more promising by extending the discussion in Section 4. In Section 5.1, the possibility of

improving the diversity of the problem is presented as an extension of Section 4.1. This will allow the use of RL for sector coupling problems, which enable decarburization in several sectors. It is interesting to check whether there is any possibility to extend the use of RL in the energy sector where the energy management problem becomes a part of a bigger problem. Section 5.2 attempts to predict such extended use of RL. The extended RL applications introduce many problems. The possibility of using the present development in RL, optimization methods, and computational facilities is discussed in Section 5.3.

5.1. Potential to extend the model concerning more diverse problems

RL have been effectively used for a set of problems related to energy systems. Reasonable progress has been achieved in most of the cases when the basic RL methods are used, indicating the potential for a notable improvement. In Section 4.1.5, it is demonstrated that RL problems can be effectively used for more diverse applications than simple dispatch or BEMS problems. Such an interlink, which connects several sectors, is also known as sector coupling [168]. Sector coupling is regarded as a potential method to decarbonize multiple sectors, such as building, transportation, and manufacturing. RL can be effectively used in such applications to consider the uncertainties and complex interrelations between different sectors whose modeling is complex when the white box approach is applied especially with the intervention of cyber-physical systems. The optimal control of resources within interconnected systems introduces several problems. First, from the training perspective, its application will be more exigent because it formulates a complex optimization problem. Second, each sector has its own interest, which may be completely different from those of other sectors. For example, the grid operation, dispatch problem, and BEMS may focus on grid stability, minimization of generation cost, and comfort of occupants, respectively. In certain cases, these objectives may be conflicting, making it difficult to minimize them simultaneously. Information sharing with different sectors is another problem. The BEMS requires sensitive information, such as the presence of people and use of equipment at specific time intervals. These are difficult to share publicly, making exigent to link the BEMS with the dispatch problem in certain cases. The formulation and use innovative methods that can handle data privacy would be interesting in this regard. Finally, there are difficulties in synchronizing these sectors mainly because of the mismatch in response time. For example, the dispatch and BEMS problems often operate within 15 min-1h time resolutions, whereas grid operation requires considerably finer time resolutions (seconds or even finer temporal resolutions). Resolving these mismatches is important when controlling the energy flows within interconnected systems. Considering all the foregoing, it can be concluded that RL has the potential to be used in solving sector-coupling problems despite the presence of several problems, some of which require the use of more advanced machine learning and optimization techniques, as discussed in detail in Sections 5.3-5.4.

5.2. RL applications beyond energy flow control

The state of the art clearly demonstrates that RL can be effectively used for a number of control problems in the energy system domain. More importantly, its potential for resolving complex problems, such as those in sector coupling, is demonstrated. This can considerably aid in energy transition and climate change mitigation. It would be interesting to investigate the potential of RL beyond simply controlling energy flows. Although RL has been effectively used with supervised and unsupervised learnings in other sectors, limited examples are found in the energy system domain.

The energy system design problem is usually linked with the optimal control problem and conducted as a bi-level optimization problem where the system operation (optimal control problem) and system-sizing problems are considered at the primary and secondary levels, respectively. Such a bi-level design approach is usually applied in microgrid design and energy systems in automobiles. Initially, simple rule-based techniques have been applied to represent the control strategy. Subsequently, fuzzy logic and MPC are introduced to facilitate more complex energy flows. However, there are a number of limitations in both gray and white box approaches, especially when considering both cyber and physical interactions in the energy system domain. RL can be an attractive alternative in this regard as it can effectively handle such complex environments because of the model-free approach [32]. This will require a substantial extension of optimization because of the mismatch in the optimization techniques employed between energy system and RL. Energy system design is usually achieved using either linear/mixed integer linear or heuristic methods, whereas RL algorithms are trained using gradient descent methods. Accordingly, linking these two problems will significantly increase the use of RL in the energy system domain. Furthermore, the possibility of utilizing RL with unsupervised learning techniques, such as clustering, will aid in more effectively locating the energy systems compared to existing approaches that only use unsupervised learning. Considering all the foregoing, it can be concluded that RL may employed beyond energy flow control, which would be an interesting new research area.

5.3. Novel trends in RL

5.3.1. Limitations of present state-of-the-art

There are many limitations in the present state of the art reinforcement learning techniques. As carefully investigated in Ref. [75], deep RL algorithms are susceptible to the choice of hyperparameter values, network architecture, reward shaping, and implementation codebase. The instability of Deep RL algorithms (the learning process exhibits high variance, and a near-optimal policy turns arbitrarily bad) tends to affect their performance adversely. Thus, in recent years, the RL research community focuses on developing stable RL algorithms with reduced variance [169,170]. Furthermore, majority of the deep RL methods fail to perform well when there is some difference between training and testing scenarios, thereby posting serious safety and security concerns. To this end, learning policies that are robust to environmental shifts, mismatched configurations, and even mismatched control actions are becoming increasingly more important. There are works that build on the robust MDP framework [171], for example [172], and [173], whereas some other leverages on the equivalence between action-robust and robust MDPs introduced by Ref. [174], for example [175], and [176]. Despite the impressive empirical progress, the robust RL objectives' training remains an open and critical challenge.

The successful application of deep RL methods is attributed to the implementation of meaningful representation learning via deep neural networks. However, for real-world problems, such as energy system optimization, standard representation techniques for a vision-based application may not be advantageous. It is therefore necessary to contextualize the representations along with deep RL methods to enhance their applicability in the energy system domain. Most current deep RL methods augment their main objective with additional losses, typically facilitating and regularizing the representation learning process [177–180]. Recently, deep RL researchers have started exploring unsupervised representation learning methods for RL [181–183].

5.3.2. Present state of the art developments: optimization in RL

In RL community, there is a recent interest to understand the standard (policy gradient-based) RL algorithms from the optimization perspective. The research community has also attempted to study the theoretical convergence properties of PG methods from a non-convex optimization perspective [184,185]. In particular, by leveraging the recent advancements in non-convex optimization, new algorithmic solutions are also being proposed for RL. Furthermore, by exploiting the minimax duality of Bellman equations, a class of stochastic primal-dual (SPD) methods that is computational and sample-efficient has been

proposed for RL [186]. The SPD Q-learning in Ref. [187] extends it to the Q-learning framework with off-policy learning. It is presumed that the energy system domain will benefit from extending the optimization-based RL with time.

Energy system operation is often considered as a multi objective optimization problem. Often conflicting objectives such as cost, comfort level, environmental impact are considered. Both Pareto and weighted objective functions are used on this regards. Within the RL community, multi-task RL is used to handle this task. Multi-task RL is a promising approach to alleviate the sample complexity issue in RL algorithms that learn individual tasks from scratch. Multi-task RL methods share structure across multiple tasks to enable more efficient learning [188,189]. Multi-task RL methods pose significant optimization challenges compared to standard RL methods that learn tasks independently from scratch [190]. By assuming additional structure on the similarity between the tasks, people have proposed more efficient optimization algorithms for Multi-task RL with provable guarantees [191]. Multi-task RL also suffers from scalability issues when the number of tasks grows large. The decentralized optimization remedies the scalability issues by distributing computation across multiple units; recently [192], have proposed scalable Multi-task RL algorithms with improved convergence guarantees.

The ever-increasing complexity of energy systems demands for reasonable improvements in the optimization techniques used for RL. As a result, there is a recent surge in customizing the online (convex) optimization algorithms for energy system-related problems [193–195]. Based on the min-max RL formulation, reductions between RL methods and online learning algorithms have been established [196], leading to the systematic development of new RL algorithms. In Ref. [197], the connections between DP and (constrained) convex optimization are established and several policy iteration algorithms in the optimization language are formulated. For example, they link conservative policy iteration to the Frank-Wolfe algorithm, mirror descent modified policy iteration to mirror descent, and Politex (policy iteration using expert prediction) to dual averaging. Accordingly, it is assumed that a considerable effort can be devoted in this direction by leveraging these relationships for more complex energy problems [196,197].

5.3.3. Distributed RL

As discussed in Sections 5.1 and 5.2, MARL can perform a major function in understanding the interactions among multiple sectors in the energy system domain. However, learning in a multi-agent environment is considerably more exigent when compared to a single-agent setting because the agent has to interact with both the environment and other agents [198]. For a recent survey on deep MARL, the reader is referred to Refs. [199,200], and for the distributed optimization-based MARL, cf [201,202]. Despite several practical problems (e.g., non-stationarity issues [203], curse of dimensionality, and computational demand) associated with deep MARL, in several recent works, empirical success in complex multi-agent scenarios has been reported. This success is achieved by carefully scaling the algorithms originally introduced to RL and multi-agent learning to deep MARL. Most of these studies, which directly use single-agent algorithms in the multi-agent setting (i.e., independent learners), lack theoretical/convergence guarantees. In some studies [204,205], the focus is mainly set on analyzing and evaluating the DRL algorithms in a multi-agent environment under cooperative, competitive, and mixed scenarios. Littman [206] studied the convergence properties of joint-action RL agents in Markov games. Recently, several MARL algorithms, such as the distributed TD learning [207], distributed Q-learning [208], and distributed actor-critic algorithm [209], have been proposed by leveraging the core idea of averaging consensus-based distributed optimization [210]. These methods achieve global consensus on optimal policy only through local computation and communication with neighboring agents. The extensions of these distributed optimization methods to recent deep RL algorithms are currently under investigation. Compared to classical RL methods, optimization-based RL methods can more effectively handle different multiple objectives that arise in real-world control problems, e.g., energy systems in the form of constraints.

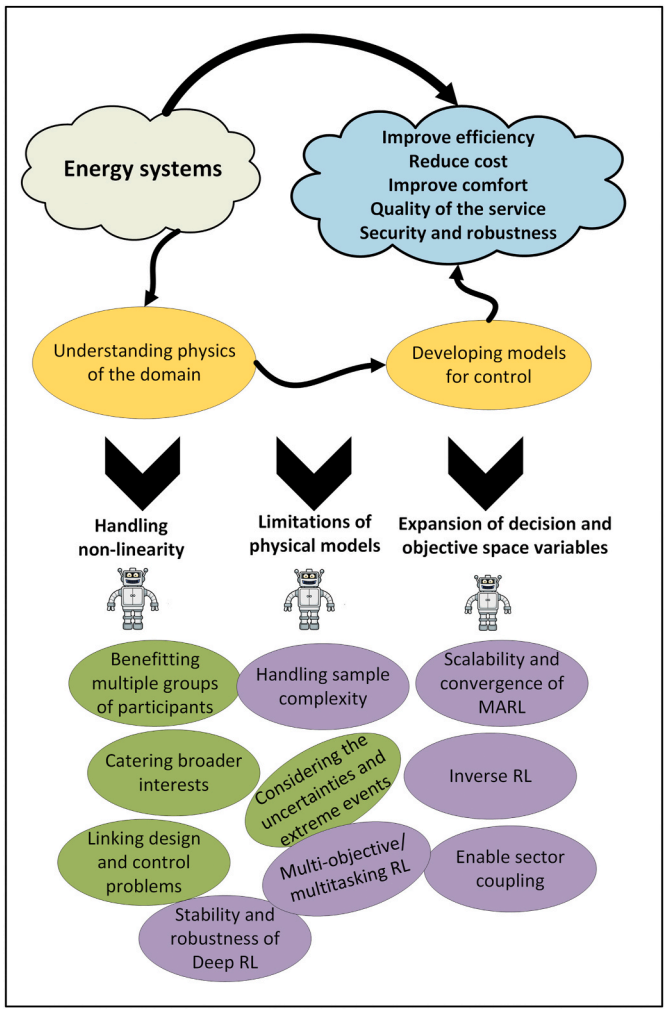

Fig. 11. State of the art and future perspectives concerning RL and energy system control.

6. Conclusions

The energy system goes through a significant transition with the decarbonization of the energy sector, notably increasing the complexity of energy systems. Integration of non-dispatchable renewable energy technologies and distributed energy storage, the introduction of complex market mechanisms, uncertainties brought by energy markets, climate and occupancy and the expansion of the energy system boundaries concerning sector coupling, etc. Demands for a paradigm shift in the methods used to control the energy systems. These demand to shift from the present state of the art method such as model-based, gray box, and rule-based methods to data-driven approaches. Three major bottlenecks are expected to be addressed by data-driven models, i.e.,

- Handling the complexity of the models used (non-linear and non-convex nature of objective functions)
- Limitations in the physical models, especially in the building sector
- Limitations in handling large state and control variable space.

Addressing these limitations will notably improve efficiency while reducing the operation cost, sustainability, user comfort, quality of service, security, and quality of the energy systems' service.

Reinforcement Learning becomes quite an attractive alternative in this regard. As a branch of machine learning, Reinforcement Learning is gradually gaining popularity beyond the machine learning community because of its broader applicability. Compared to model predictive control strategies, Reinforcement Learning employs a model-free approach and does not require convergence guarantees, thereby significantly increasing its applicability. Also, the capability to handle a large decision space with a little knowledge about the problem physics make it competitive compared with the other rule-based controllers used in the present state of the art (cf. Fig. 11). Therefore, a progressive improvement is observed in publications that report the use of Reinforcement Learning.

The publications within the energy system domain grow rapidly, focusing on a broad group of problems, making it difficult to consider one by one. A novel approach is proposed in the present study to address this problem by using a top-down approach. In this study, the publications that discuss the resolution of energy system-related problems using Reinforcement Learning are clustered into seven groups. Six groups are related to specific control problems, such as building energy management, dispatch, energy systems in hybrid vehicles, energy markets, grid, and energy devices. Within this classification, the study investigates.

1. The complexity of the problem within each problem class
   (a) Handling non-linearity
   (b) Use of approximation models/data-driven models
   (c) Expansion of decision and objective space variables
2. Types of RL methods used
3. How far the present state of art methods has been successful?

The study reveals that most RL (45%) studies focus on either building energy management or dispatch problem. Most of the papers related to building energy management systems focus on HVAC, whereas a few publications center on lighting and control of blinds together with the HVAC. More diversity is observed in shifting to the dispatch problem. Energy systems with considerably complex system configurations are considered in the dispatch problem while focusing on the uncertainties in renewable generation, demand, and price signals in the grid. The study reveals that the dispatching problem has a close link between the energy market and the grid in addition to the building energy management systems. Multi-agent Reinforcement Learning has been used to couple the dispatch and energy markets; this enables to schedule distributed energy systems considering day-ahead and spot markets. Grid models have been used along with Reinforcement Learning algorithms to guarantee the secure and stable operation of the distribution networks while accommodating distributed energy sources. The study reveals the potential of linking dispatch problems with the optimal power flow problem in the grid; this will facilitate to improve the efficiency of the grid while guaranteeing stability. The flexibility demonstrated by the dispatch problem to link with building energy management, grid, and energy markets makes it an ideal candidate to act as the central hub while maintaining interactions between many participants within the energy system domain. However, the connectivity of energy devices and vehicle energy systems between the other sectors are difficult to observe. These sectors focus on finer time resolutions than in other sectors, making them more difficult to link with other sectors. Within these different domains, RL has shown the potential to consider large state spaces and complex nonlinear models that are difficult to handle using the other existing techniques.

Several major bottlenecks are observed when looking into the present state of the art. Most studies lack proper benchmarking compared to model-based approaches or gray-box models; this makes it difficult to make any conclusions regarding performance improvement. The development of a public repository of computational codes and test cases to validate performance improvement could be considerably advantageous for the research community, as evidenced by its implementation in other ML communities, such as computer vision. The reproducibility of the results is another major issue that has not been discussed broadly. Comparing the performance of RL algorithms taking case studies with different environments will be helpful in this regard. Although RL can adequately solve integrated problems in several sectors, only a limited number of publications have discussed its broad application. One of the most remarkable observations is that most studies do not use deep learning techniques; instead, the tabular method or shallow network for function approximation is employed. In practically all publications, the use of existing libraries and optimization algorithms is reported with regard to the implementation of Reinforcement Learning. The development of RL algorithms to cater to bottlenecks in the energy system control problems is not reported. Considering all of the aforementioned aspects, there remain many open questions relative to both energy systems and machine learning. The answers to such questions are expected to lead to significant improvements in the energy system domain. However, even with the current limited use of reinforcement, learning methods exhibit a 10–20% performance improvement in many applications (especially in building energy management), although there are few cases where the model predictive control outperforms the RL within a considerably close margin.

The operation of the energy systems will benefit from progress in Reinforcement Learning techniques to address many open-ended problems that have not been addressed. The study reveals that improving the participation of multiple agents having different priorities will become more common, especially with the open energy markets where Reinforcement Learning could immensely help improve the participants' profits and maintain robust operation. Given the recent advances in the robust Reinforcement Learning literature, those robust RL methods could be utilized in energy systems to handle uncertainty in the system parameters. The possibility to accommodate a large pool of participants will be beneficial, especially concerning applications like energy internet where a large group of devices having energy storage (for example, mobile phones, dishwashers, etc.) will interact with the energy systems. One of the promising directions is sector coupling. Consideration of multiple objectives within the control problem is another important aspect of the energy system domain. For example, minimizing both cost and emissions are considered vital in the dispatch problem. Introducing multi-tasking RL would be quite interesting in this regard, which can address this specific issue. The applications of Reinforcement Learning algorithms to link several domains will help improve the penetration levels of renewable energy technologies while minimizing the generation of CO_2 in these sectors; the shift into multi-agent Reinforcement Learning will be beneficial in this regard. One major limitation observed when extending the scope of the problem is the oversimplification. It is noteworthy to investigate possible ways to consider detailed physics of the problems while extending the problem's scope where linking existing physical models with Reinforcement Learning techniques would be immensely helpful. Developing a hybrid technique to link model predictive control and Reinforcement Learning will also be an interesting approach to improve energy efficiency in building energy systems. Furthermore, Reinforcement Learning can be used to locate distributed energy systems where Reinforcement Learning can be employed with clustering algorithms. Therefore, advances in Reinforcement Learning techniques will essentially play a vital role towards the improvements in the energy sector.

The study reveals that Reinforcement Learning can effectively address many limitations in the present state of the art methods used to operate energy systems. It has already presented a significant potential as a major candidate to address the energy system operation problem. More importantly, the potential of data-driven methods to address complex energy-related problems is significant. The advances taking place in the machine learning community play a vital role in this regard, which facilitates Reinforcement Learning algorithms to be used for a diverse group of problems. Based on the literature review, the authors have noted the potential of RL beyond its use in control problems.

Considering all these aspects, RL definitely has a broad scope of application with a huge potential to resolve global problems in the energy sector.

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] Transforming our world: the 2030 agenda for sustainable development. Sustainable development knowledge platform. [Online; accessed 2020-03-26].
[2] Mohajeri Nahid, Perera ATD. Silvia coccolo, Lucas Mosca, Morgane Le guen, and Jean-Louis scartezzini. Integrating urban form and distributed energy systems: assessment of sustainable development scenarios for a swiss village to 2050. Renew Energy 12 2019;143:810–26.
[3] Mauree Dasaraden, Naboni Emanuele, Coccolo Silvia, Perera ATD, Nik Vahid M, Scartezzini Jean-Louis. A review of assessment methods for the urban environment and its energy sustainability to guarantee climate adaptation of future cities. Renew Sustain Energy Rev 9 2019;112:733–46.
[4] Musonye Xavier S, Davísdóttir Brynhildur, Kristjánsson Ragnar, Eyjólfur I Ásgeirsson, Stefánsson Hlynur. Integrated energy systemsâ€™ modeling studies for sub-saharan africa: a scoping review. Renew Sustain Energy Rev 2020;128:109915.
[5] Manfren Massimiliano, Caputo Paola, Costa Gaia. Paradigm shift in urban energy systems through distributed generation: methods and models. Appl Energy 4 2011;88(4):1032–48.
[6] Ackermann Thomas, Andersson Göran, Söder Lennart. Distributed generation: a definition1in addition to this paper, a working paper entitled 'distributed power generation in a deregulated market environment' is available. the aim of this working paper is to start a discussion regarding different aspects of distributed generation. this working paper can be obtained from one of the authors, thomas ackermann.1. Elec Power Syst Res 4 2001;57(3):195–204.
[7] Perera ATD, Wickremasinghe DMIJ, Mahindarathna DVS, Attalage RA, Perera KKCK, Bartholameuz EM. Sensitivity of internal combustion generator capacity in standalone hybrid energy systems. Energy 3 2012;39(1):403–11.
[8] Sansavini G, Piccinelli R, Golea LR, Zio E. A stochastic framework for uncertainty analysis in electric power transmission systems with wind generation. Renew Energy 2014;64(71–81):4.
[9] Zio E, Sansavini G. Vulnerability of smart grids with variable generation and consumption: a system of systems perspective. IEEE Trans Syst Man, and Cybernet: Systems 5 2013;43(3):477–87.
[10] Liu Teng, Yuan Zou, Liu Dexing, Sun Fengchun. Reinforcement learning based energy management strategy for a hybrid electric tracked vehicle. Energies 7 2015;8(7):7243–60.
[11] James Keirstead, Jennings Mark, Sivakumar Aruna. A review of urban energy system models: approaches, challenges and opportunities. Renew Sustain Energy Rev 8 2012;16(6):3847–66.
[12] Peter Palensky, Widl Edmund, Elsheikh Atiyah. Simulating cyber-physical energy systems: challenges, tools and methods. IEEE Trans Syst Man, and Cybernet: Systems 2013;44(3):318–26.
[13] Derler P, Lee EA, Sangiovanni Vincentelli A. Modeling cyber–physical systems. Proc IEEE 2012;100(1):13–28. 1.
[14] Sridhar S, Hahn A, Govindarasu M. Cyber–physical system security for the electric power grid. Proc IEEE 1 2012;100(1):210–24.
[15] Mavromatidis Georgios, Orehounig Kristina, Jan Carmeliet. A review of uncertainty characterisation approaches for the optimal design of distributed energy systems. Renew Sustain Energy Rev 5 2018;88(258–277).
[16] Perera ATD, Nik Vahid M, Chen Deliang, Scartezzini Jean-Louis, Hong Tianzhen. Quantifying the impacts of climate change and extreme climate events on energy systems. Nat Energy 2 2020;5(2):150–9.
[17] Yang Li, Wang Chunling, Li Guoqing, Wang Jinlong, Zhao Dongbo, Chen Chen. Improving operational flexibility of integrated energy system with uncertain renewable generations considering thermal inertia of buildings. Energy Convers Manag 2020;207:112526.
[18] Panteli Mathaios, Mancarella Pierluigi. Influence of extreme weather and climate change on the resilience of power systems: impacts and possible mitigation strategies. Elec Power Syst Res 10 2015;127:259–70.
[19] LeCun Yann, Bengio Yoshua, Hinton Geoffrey. Deep learning. Nature 5 2015;521(7553):436–44. number: 7553 publisher: Nature Publishing Group.
[20] Soteris A, Kalogirou. Applications of artificial neural-networks for energy systems. Appl Energy 2000;67(1):17–35. 9.
[21] Bishop Christopher M. Pattern recognition and machine learning. springer; 2006.
[22] Murphy Kevin P. Machine learning: a probabilistic perspective. MIT press; 2012.
[23] Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Rusu Andrei A, veness Joel, Bellemare Marc G, Graves Alex, Riedmiller Martin, Fidjeland Andreas K, Ostrovski Georg, Petersen Stig, Beattie Charles, Sadik Amir, Antonoglou Ioannis, King Helen, Kumaran Dharshan, Wierstra Daan, Legg Shane, Hassabis Demis. Human-level control through deep reinforcement learning. Nature 2 2015;518(7540):529–33.
[24] Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Graves Alex, Antonoglou Ioannis, Wierstra Daan, Riedmiller Martin. Playing atari with deep reinforcement learningvol. 12; 2013. arXiv:1312.5602 [cs], arXiv: 1312.5602.
[25] Cheng Lefeng, Tao Yu. A new generation of ai: a review and perspective on machine learning technologies applied to smart energy and electric power systems. Int J Energy Res 2019;43(6):1928–73.
[26] Han Mengjie, May Ross, Zhang Xingxing, Wang Xinru, Pan Song, Yan Da, Jin Yuan, Xu Liguo. A review of reinforcement learning methodologies for controlling occupant comfort in buildings. Sustain Cities and Soc 2019;51(101748):11.
[27] Vázquez-Canteli José R, Nagy Zoltán. Reinforcement learning for demand response: a review of algorithms and modeling techniques. Appl Energy 2 2019;235:1072–89.
[28] Wang Zhe, Hong Tianzhen. Reinforcement learning for building controls: the opportunities and challenges. Appl Energy 2020;269:115036.
[29] Hou Zhong-Sheng, Wang Zhuo. From model-based control to data-driven control: survey, classification and perspective. Inf Sci 2013;235(3–35):6.
[30] Ernst D, Glavic M, Capitanescu F, Wehenkel L. Reinforcement learning versus model predictive control: a comparison on a power system problem. IEEE Trans. Syst. Man Cybernet. Part B (Cybernetics) 2009;39(2):517–29. 4.
[31] Perera ATD, Nik Vahid M, Mauree Dasaraden, Scartezzini Jean-Louis. An integrated approach to design site specific distributed electrical hubs combining optimization, multi-criterion assessment and decision making. Energy 2017;134(103–120). 9.
[32] Perera ATD, Wickramasinghe PU, Nik Vahid M, Scartezzini Jean-Louis. Introducing reinforcement learning to the energy system design process. Appl Energy 2020;262:114580. 3.
[33] Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Rusu Andrei A, Veness Joel, Bellemare Marc G, Graves Alex, Riedmiller Martin, Andreas K Fidjeland, Ostrovski Georg, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529.
[34] Silver David, Schrittwieser Julian, Simonyan Karen, Antonoglou Ioannis, Huang Aja, Guez Arthur, Hubert Thomas, Baker Lucas, Lai Matthew, Bolton Adrian, et al. Mastering the game of go without human knowledge. Nature 2017;550(7676):354.
[35] Lillicrap Timothy P, Hunt Jonathan J, Alexander Pritzel, Heess Nicolas, Tom Erez, Tassa Yuval, Silver David, Wierstra Daan. Continuous control with deep reinforcement learning. 2015. arXiv preprint arXiv:1509.02971.
[36] Levine Sergey, Finn Chelsea, Trevor Darrell, Abbeel Pieter. End-to-end training of deep visuomotor policies. J Mach Learn Res 2016;17(1):1334–73.
[37] Sutton Richard S, Barto Andrew G. Reinforcement learning: an introduction. MIT press; 2018.
[38] Powell Warren B. Approximate dynamic programming: solving the curses of dimensionality, ume 703. John Wiley & Sons; 2007.
[39] Bertsekas Dimitri P, Tsitsiklis John N. Neuro-dynamic programming. MA: Athena Scientific Belmont; 1996.
[40] Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons; 2014.
[41] Silver David. Ucl course on rl. Accessed: 2020-04-07, https://www.davidsilver.uk/teaching/; 2015.
[42] Sutton Richard S. Dyna, an integrated architecture for learning, planning, and reacting. ACM SIGART Bull 1991;2(4):160–3.
[43] Peng Baolin, Li Xiujun, Gao Jianfeng, Liu Jingjing, Wong Kam-Fai, Su Shang-Yu. Deep dyna-q: integrating planning for task-completion dialogue policy learning. 2018. arXiv preprint arXiv:1801.06176.
[44] Amir-massoud Farahmand, Barreto Andre, Nikovski Daniel. Value-aware loss function for model-based reinforcement learning. Artificial intelligence and statistics. 2017. p. 1486–94.
[45] Wang Xin, Dietterich Thomas G. Model-based policy gradient reinforcement learning. Proceedings of the 20th international conference on machine learning. ICML-03); 2003. p. 776–83.
[46] Abachi Romina, Ghavamzadeh Mohammad, Farahmand Amir-massoud. Policy-aware model learning for policy gradient methods. 2020. arXiv preprint arXiv:2003.00030.
[47] Kudashkina Katya, Chockalingam Valliappa, Taylor Graham W, Bowling Michael. Sample-efficient model-based actor-critic for an interactive dialogue task. 2020. arXiv preprint arXiv:2004.13657.
[48] Clavera Ignasi, Fu Violet, Abbeel Pieter. Model-augmented actor-critic: backpropagating through paths. 2020. arXiv preprint arXiv:2005.08068.
[49] Zhang Shangtong, Boehmer Wendelin, Whiteson Shimon. Deep residual reinforcement learning. 2019. arXiv preprint arXiv:1905.01072.
[50] Gavin A. Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, ume 37. England: University of Cambridge, Department of Engineering Cambridge; 1994.
[51] Watkins Christopher JC H, Dayan Peter. Q-learning. Mach learn 1992;8(3–4):279–92.
[52] Van Hasselt Hado, Guez Arthur, Silver David. Deep reinforcement learning with double q-learning. Thirtieth AAAI conference on artificial intelligence. 2016.
[53] Wang Ziyu, Tom Schaul, Hessel Matteo, Van Hasselt Hado, Lanctot Marc, De Freitas Nando. Dueling network architectures for deep reinforcement learning. 2015. arXiv preprint arXiv:1511.06581.
[54] John Schulman, Levine Sergey, Abbeel Pieter, Jordan Michael, Moritz Philipp. Trust region policy optimization. International conference on machine learning. 1889–1897. p. 2015.
[55] Silver David, Guy Lever, Heess Nicolas, Degris Thomas, Wierstra Daan, Riedmiller Martin. Deterministic policy gradient algorithms. ICML; 2014.

[56] Scott Fujimoto, Herke van Hoof, Meger David. Addressing function approximation error in actor-critic methods. 2018. arXiv preprint arXiv: 1802.09477.

[57] Haarnoja Tuomas, Zhou Aurick, Abbeel Pieter, Levine Sergey. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. 2018. arXiv preprint arXiv:1801.01290.

[58] Weng Lilian. Policy gradient algorithms. lilianweng.github.io/lil-log; 2018.

[59] Sun Wen, Jiang Nan, Krishnamurthy Akshay, Agarwal Alekh, Langford John. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. Conference on learning theory. 2019. p. 2898–933.

[60] Tu Stephen, Recht Benjamin. The gap between model-based and model-free methods on the linear quadratic regulator: an asymptotic viewpoint. Conference on learning theory. 2019. p. 3036–83.

[61] Kurutach Thanard, Clavera Ignasi, Duan Yan, Aviv Tamar, Abbeel Pieter. Model-ensemble trust-region policy optimization. 2018. arXiv preprint arXiv: 1802.10592.

[62] Luo Yuping, Xu Huazhe, Li Yuanzhi, Tian Yuandong, Trevor Darrell, Ma Tengyu. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. International conference on learning representations. 2019.

[63] Wang Tingwu, Bao Xuchan, Clavera Ignasi, Hoang Jerrick, Wen Yeming, Langlois Eric, Zhang Shunshi, Zhang Guodong, Abbeel Pieter, Jimmy Ba. Benchmarking model-based reinforcement learning. 2019. arXiv preprint arXiv: 1907.02057.

[64] Lange Sascha, Gabel Thomas, Riedmiller Martin. Batch reinforcement learning. Reinforcement learning. Springer; 2012. p. 45–73.

[65] Lagoudakis Michail G, Parr Ronald. Least-squares policy iteration. J Mach Learn Res 2003;4(Dec):1107–49.

[66] Ernst Damien, Geurts Pierre, Wehenkel Louis. Tree-based batch mode reinforcement learning. J Mach Learn Res 2005;6(Apr):503–56.

[67] Martin Riedmiller. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. European conference on machine learning. Springer; 2005. p. 317–28.

[68] Agarwal Rishabh, Dale Schuurmans, Norouzi Mohammad. Striving for simplicity in off-policy deep reinforcement learning. 2019. arXiv preprint arXiv: 1907.04543.

[69] Scott Fujimoto, Meger David, Precup Doina. Off-policy deep reinforcement learning without exploration. 2018. arXiv preprint arXiv:1812.02900.

[70] Kumar Aviral, Fu Justin, Soh Matthew, Tucker George, Levine Sergey. Stabilizing off-policy q-learning via bootstrapping error reduction. Advances in neural information processing systems. 2019. p. 11761–71.

[71] Dhariwal Prafulla, Hesse Christopher, Klimov Oleg, Nichol Alex, Plappert Matthias, Radford Alec, John Schulman, Sidor Szymon, Wu Yuhuai, Zhokhov Peter. Openai baselines. 2017.

[72] Hill Ashley, Raffin Antonin, Ernestus Maximilian, Adam Gleave, Kanervisto Anssi, Traore Rene, Dhariwal Prafulla, Hesse Christopher, Klimov Oleg, Nichol Alex, Plappert Matthias, Radford Alec, John Schulman, Sidor Szymon, Wu Yuhuai. Stable baselines. https://github.com/hill-a/stable-baselines; 2018.

[73] Alexander Kuhnle, Schaarschmidt Michael, Fricke Kai. Tensorforce: a tensorflow library for applied reinforcement learning. Web page; 2017.

[74] D'Eramo Carlo, Tateo Davide, Bonarini Andrea, Restelli Marcello, Peters Jan. Mushroomrl: simplifying reinforcement learning research. https://github.com/MushroomRL/mushroom-rl; 2020.

[75] Henderson Peter, Islam Riashat, Bachman Philip, Pineau Joelle, Precup Doina, Meger David. Deep reinforcement learning that matters. 2017. arXiv preprint arXiv:1709.06560.

[76] Park June Young, Dougherty Thomas, Fritz Hagen, Nagy Zoltan. Lightlearn: an adaptive and occupant centered controller for lighting based on reinforcement learning. Build Environ 1 2019;147:397–414.

[77] Cheng Zhijin, Zhao Qianchuan, Wang Fulin, Jiang Yi, Xia Li, Ding Jinlei. Satisfaction based q-learning for integrated lighting and blind control. Energy Build 2016;127(43–55):9.

[78] Eller Lukas, Siafara Lydia C, Sauter Thilo. Adaptive control for building energy management using reinforcement learning. IEEE international conference on industrial technology (ICIT), 2 2018. 2018. p. 1562–7 [ISSN: null].

[79] Valladares William, Galindo Marco, Gutiérrez Jorge, Wu Wu-Chieh, Liao Kuo-Kai, Liao Jen-Chung, Lu Kuang-Chin, Wang Chi-Chuan. Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. Build Environ 5 2019;155(105–117).

[80] Heo SungKu, Nam KiJeon, Loy-Benitez Jorge, Qian Li, Lee SeungChul, Yoo ChangKyoo. A deep reinforcement learning-based autonomous ventilation control system for smart indoor air quality management in a subway station. Energy Build 2019;202(109440):11.

[81] Zheng Wen, O'Neill Daniel, Hamid Maei. Optimal demand response using device-based reinforcement learning. IEEE Trans Smart Grid 2015;6(5):2312–24. 9.

[82] Bollenbacher J, Rhein B. Optimal configuration and control strategy in a multi-carrier-energy system using reinforcement learning methods. International energy and sustainability conference (IESC), 10 2017. 2017. p. 1–6 [ISSN: null].

[83] Sun Qiuye, Wang Danlu, Ma Dazhong, Huang Bonan. Multi-objective energy management for we-energy in energy internet using reinforcement learning. IEEE symposium series on computational intelligence (SSCI), 11 2017, vols. 1–6; 2017 [ISSN: null].

[84] Mbuwir Brida V, Kaffash Mahtab, Deconinck Geert. Battery scheduling in a residential multi-carrier energy system using reinforcement learning. IEEE international conference on communications, control, and computing technologies for smart grids. SmartGridComm); 2018. p. 1–6. 2018.

[85] Kofinas Panagiotis, George Vouros, Anastasios I Dounis. Energy management in solar microgrid via reinforcement learning using fuzzy reward. Adv Build Energy Res 2018;12(1):97–115. 1.

[86] Kofinas P, Dounis AI, Vouros GA. Fuzzy q-learning for multi-agent decentralized energy management in microgrids. Appl Energy 2018;219(53–67):6.

[87] Foruzan Elham, , Leen-Kiat Soh, Asgarpoor Sohrab. Reinforcement learning approach for optimal distributed energy management in a microgrid. IEEE Trans Power Syst 9 2018;33(5):5749–58.

[88] Zarrabian Sina, Belkacemi Rabie, Adeniyi A. Babalola. Reinforcement learning approach for congestion management and cascading failure prevention with experimental application. Elec Power Syst Res 2016;141:179–90. 12.

[89] Rocchetta R, Bellani L, Compare M, Zio E, Patelli E. A reinforcement learning framework for optimal operation and maintenance of power grids. Appl Energy 5 2019;241:291–301.

[90] Ermon Stefano, Xue Yexiang, Gomes Carla, Selman Bart. Learning policies for battery usage optimization in electric vehicles. Mach Learn 7 2013;92(1):177–94.

[91] Xiong Rui, Cao Jiayi, Yu Quanqing. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. Appl Energy 2 2018;211:538–48.

[92] Namireddy Praveen Reddy, Pasdeloup David, Karbalaye Zadeh Mehdi, Roger Skjetne. An intelligent power and energy management system for fuel cell/battery hybrid electric vehicle using reinforcement learning. IEEE transportation electrification conference and expo (ITEC), 6 2019, vols. 1–6. ISSN; 2019. p. 2377–5483.

[93] James Brusey, Hintea Diana, Gaura Elena, Beloe Neil. Reinforcement learning-based thermal comfort control for vehicle cabins. Mechatronics 2018;50:413–21. 4.

[94] Claessens Bert J, Vanhoudt D, Desmedt J, Ruelens F. Model-free control of thermostatically controlled loads connected to a district heating network. Energy Build 2018;159(1–10):1.

[95] Lee Sangyoon, Choi Dae-Hyun. Reinforcement learning-based energy management of smart home with rooftop solar photovoltaic system, energy storage system, and home appliances. Sensors 2019;19(18):9. PMID: 31547320 PMCID: PMC6767655.

[96] Chen Yujiao, Norford Leslie K, Samuelson Holly W, Ali Malkawi. Optimal control of hvac and window systems for natural ventilation through reinforcement learning. Energy Build 6 2018;169:195–205.

[97] Peirelinck Thijs, Ruelens Frederik, Decnoninck Geert. Using reinforcement learning for optimizing heat pump control in a building model in modelica. IEEE international energy conference (ENERGYCON), 6 2018. 2018. p. 1–6 [ISSN: null].

[98] Kazmi Hussain, Suykens Johan, Balint Attila, Driesen Johan. Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads. Appl Energy 2019;238:1022–35. 3.

[99] Zhang Zhiang, Chong Adrian, Pan Yuqi, Zhang Chenlu, Lam Khee Poh. Whole building energy model for hvac optimal control: a practical framework based on deep reinforcement learning. Energy Build 2019;199:472–90. 9.

[100] Liu Tao, Xu Chengliang, Guo Yabin, Chen Huanxin. A novel deep reinforcement learning based methodology for short-term hvac system energy consumption prediction. Int J Refrig 2019;107(39–51):11.

[101] Young Ran Yoon, Moon Hyeun Jun. Performance based thermal comfort control (ptcc) using deep reinforcement learning for space cooling. Energy Build 2019; 203(109420):11.

[102] Wei Tianshu, Wang Yanzhi, Qi Zhu. Deep reinforcement learning for building hvac control. 54th ACM/EDAC/IEEE design automation conference (DAC), 6 2017. 2017. p. 1–6 [ISSN: null].

[103] Yu Zhen, Arthur Dexter. Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning. Contr Eng Pract 2010;18 (5):532–9. 5.

[104] de Gracia Alvaro, Fernández Cèsar, Albert Castell, Mateu Carles, Luisa F, Cabeza. Control of a pcm ventilated facade using reinforcement learning techniques. Energy Build 2015;106(234–242):11.

[105] Wang Yuan, Velswamy Kirubakaran, Huang Biao. A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. Processes 2017;5(3):46. 9.

[106] Ruelens F, Claessens BJ, Quaiyum S, De Schutter B, Babuáka R, Belmans R. Reinforcement learning applied to an electric water heater: from theory to practice. IEEE Trans Smart Grid 7 2018;9(4):3792–800.

[107] Ruelens Frederik, Iacovella Sandro, Claessens Bert J, Belmans Ronnie. Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. 2015. p. 6. arXiv:1506.01054 [cs], arXiv: 1506.01054.

[108] Jia Ruoxi, Jin Ming, Sun Kaiyu, Hong Tianzhen, Spanos Costas. Advanced building control via deep reinforcement learning. Energy Procedia 2 2019;158: 6158–63.

[109] Kazmi H, D'Oca S, Delmastro C, Lodeweyckx S, Corgnati SP. Generalizable occupant-driven optimization model for domestic hot water production in nzeb. Appl Energy 2016;175(1–15):8.

[110] Schmidt Mischa, Victoria Moreno M, Schülke Anett, Macek Karel, Karel Mařík, Gordaliza Pastor Alfonso. Optimizing legacy building operation: the evolution into data-driven predictive cyber-physical systems. Energy Build 8 2017;148: 257–79.

[111] Kazmi Hussain, Mehmood Fahad, Lodeweyckx Stefan, Driesen Johan. Gigawatt-hour scale savings on a budget of zero: deep reinforcement learning based optimal control of hot water systems. Energy 2018;144(159–168):2.

[112] Hua Haochen, Qin Yuchao, Hao Chuantong, Cao Junwei. Optimal energy management strategies for energy internet via deep reinforcement learning approach. Appl Energy 2019;239(598–609):4.
[113] Zhou Min, Wang Bo, Li Tiantian, Watada Junzo. A data-driven approach for multi-objective unit commitment under hybrid uncertainties. Energy 2018;164:722–33. 12.
[114] Ji Ying, Wang Jianhui, Xu Jiacan, Fang Xiaoke, Zhang Huaguang. Real-time energy management of a microgrid using deep reinforcement learning. Energies 2019;12(12):2291. 1.
[115] Bharat Menon R, Menon Sangeetha B, Srinivasan Dipti, Jain Lakhmi. Online reinforcement learning in multi-agent systems for distributed energy systems. IEEE Innovative Smart Grid Technologies - asia (ISGT ASIA), 5 2014. 2014. p. 791–6. ISSN: 2378-8542.
[116] Sheikhi A, Rayati M, Ranjbar AM. Dynamic load management for a residential customer; reinforcement learning approach. Sustain Cities and Soc 2016;24(42–51):7.
[117] Yu T, Wang HZ, Zhou B, Chan KW, Tang J. Multi-agent correlated equilibrium q-learning for coordinated smart generation control of interconnected power grids. IEEE Trans Power Syst 7 2015;30(4):1669–79.
[118] Berlink Heider, Kagan Nelson, Anna Helena Reali Costa. Intelligent decision-making for smart home energy management. J Intell Rob Syst 2015;80(1):331–54. 12.
[119] Wang Yanzhi, Lin Xue, Pedram Massoud. A near-optimal model-based control algorithm for households equipped with residential photovoltaic power generation and energy storage systems. IEEE Trans Sustain Energy 1 2016;7(1):77–86.
[120] Yu Tao, Lei Xi, Yang Bo, Zhao Xu, Jiang Lin. Multiagent stochastic dynamic game for smart generation control. J Energy Eng 2016;142(1):3. 04015012.
[121] Kumar Venayagamoorthy Ganesh, Sharma Ratnesh K, Gautam Prajwal K, Ahmadi Afshin. Dynamic energy management system for a smart microgrid. IEEE Trans Neural Network Learn Syst 8 2016;27(8):1643–56.
[122] Guan Chenxiao, Wang Yanzhi, Lin Xue, Nazarian Shahin, Pedram Massoud. Reinforcement learning-based control of residential energy storage systems for electric bill minimization. 12th annual IEEE consumer communications and networking conference (CCNC), 1 2015. 2015. p. 637–42. ISSN: 2331-9860.
[123] Abdulla Khalid, De Hoog Julian, Kent Steer, Wirth Andrew, Halgamuge Saman. Multi-resolution dynamic programming for the receding horizon control of energy storage. IEEE Trans Sustain Energy 1 2019;10(1):333–43.
[124] Liu Weirong, Peng Zhuang, Liang Hao, Peng Jun, Huang Zhiwu. Distributed economic dispatch in microgrids based on cooperative reinforcement learning. IEEE Trans Neural Network Learn Syst 6 2018;29(6):2192–203.
[125] Bukenberger Jesse, Palmintier Bryan. Stochastic generation capacity expansion planning with approximate dynamic programming. 1–5. IEEE/PES Transmission and Distribution Conference and Exposition (T D); 2018. 4 2018. ISSN: 2160-8563.
[126] Niklas Ebell, Pruckner Marco. Coordinated multi-agent reinforcement learning for swarm battery control. IEEE Canadian conference on electrical computer engineering (CCECE), 5 2018. 2018. p. 1–4. ISSN: 2576-7046.
[127] Qiu Xin, Nguyen Tu A, Crow Mariesa L. Heterogeneous energy storage optimization for microgrids. IEEE Trans Smart Grid 2016;7(3):1453–61. 5.
[128] Hassan Shirzeh, Naghdy Fazel, Ciufo Philip, Ros Montserrat. Balancing energy in the smart grid using distributed value function (dvf). IEEE Trans Smart Grid 2015;6(2):808–18. 3.
[129] Naghibi-Sistani MB, Akbarzadeh-Tootoonchi MR, Javidi-Dashte Bayaz MH, Rajabi-Mashhadi H. Application of q-learning with temperature variation for bidding strategies in market based power systems. Energy Convers Manag 2006;47(11):1529–38. 7.
[130] Nanduri Vishnu, Kazemzadeh Narges. Economic impact assessment and operational decision making in emission and transmission constrained electricity markets. Appl Energy 2012;96(212–221):8.
[131] Chen Tao, Su Wencong. Local energy trading behavior modeling with deep reinforcement learning. IEEE Access 2018;6:62806–14.
[132] Lu Renzhi, Ho Hong Seung, Zhang Xiongfeng. A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach. Appl Energy 2018;220:220–30. 6.
[133] Lu Renzhi, Ho Hong Seung. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. Appl Energy 2 2019;236:937–49.
[134] Rahimi-Kian A, Tabarraei H, Sadeghi B. Reinforcement learning based supplier-agents for electricity markets. Proceedings of the 2005 IEEE international symposium on, mediterrean conference on control and automation intelligent control. 2005. p. 1405–10. 6 2005. ISSN: 2158-9879.
[135] Igushi Kenta, Takaya Ogiso, Yamauchi Koichiro. Acceleration of reinforcement learning via game-based renewal energy management system. Joint 7th international conference on soft computing and intelligent systems (SCIS) and 15th international symposium on advanced intelligent systems (ISIS). 2014. p. 415–20. 2014.
[136] Kim Byung-Gook, Zhang Yu, van der Schaar Mihaela, Lee Jang-Won. Dynamic pricing and energy consumption scheduling with reinforcement learning. IEEE Trans Smart Grid 2016;7(5):2187–98. 9.
[137] Peters Markus, Ketter Wolfgang, Saar-Tsechansky Maytal, Collins John. A reinforcement learning approach to autonomous decision-making in smart electricity markets. Mach Learn 2013;92(1):7. 5–39.
[138] Carvalho Marco, Perez Carlos, Adrian Granados. An adaptive multi-agent-based approach to smart grids control and optimization. Energy Syst 2012;3(1):61–76. 3.
[139] Mahmoud MS, Abouheaf M, Sharaf A. Reinforcement learning control approach for autonomous microgrids. Int J Model Simulat 2019;(1–10):8. 0(0).
[140] Cao Junwei, Zhang Wanlu, Xiao Zeqing, Hua Haochen. Reactive power optimization for transient voltage stability in energy internet via deep reinforcement learning approach. Energies 2019;12(8):1556. 1.
[141] Sun Jian, Zhu Zhiqin, Li Huaqing, Chai Yi, Qi Guanqiu, Wang Huiwei, Yu Hen Hu. An integrated critic-actor neural network for reinforcement learning with application of ders control in grid frequency regulation. Int J Electr Power Energy Syst 2019;111:286–99. 10.
[142] Liu Teng, Yuan Zou, Liu Dexing, Sun Fengchun. Reinforcement learning of adaptive energy management with transition probability for a hybrid electric tracked vehicle. IEEE Trans Ind Electron 2015;62(12):7837–46. 12.
[143] Liu Teng, Wang Bo, Yang Chenglang. Online Markov chain-based energy management for a hybrid tracked vehicle with speedy q-learning. Energy 2018;160:544–55. 10.
[144] Yuan Zou, Liu Teng, Liu Dexing, Sun Fengchun. Reinforcement learning-based real-time energy management for a hybrid tracked vehicle. Appl Energy 2016;171:372–82. 6.
[145] Yuan Jingni, Yang Lin, Chen Qu. Intelligent energy management strategy based on hierarchical approximate global optimization for plug-in fuel cell hybrid electric vehicles. Int J Hydrogen Energy 2018;43(16):8063–78. 4.
[146] Wu Yuankai, Tan Huachun, Peng Jiankun, Zhang Hailong, He Hongwen. Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus. Appl Energy 2019;247:454–66. 8.
[147] Zhou Quan, Li Ji, Shuai Bin, Williams Huw, He Yinglong, Li Ziyang, Xu Hongming, Yan Fuwu. Multi-step reinforcement learning for model-free predictive energy management of an electrified off-highway vehicle. Appl Energy 2019;255:113755. 12.
[148] Qi Xuewei, Wu Guoyuan, Boriboonsomsin Kanok, Barth Matthew J, Gonder Jeffrey. Data-driven reinforcement learningâ€“based real-time energy management system for plug-in hybrid electric vehicles. Transport Res Rec 2016;2572(1):1–8. 1.
[149] Kofinas P, Doltsinis S, Dounis AI, Vouros GA. A reinforcement learning approach for mppt control method of photovoltaic sources. Renew Energy 2017;108:461–73. 8.
[150] Anderlini Enrico, Forehand David IM, Bannon Elva, Abusara Mohammad. Control of a realistic wave energy converter model using least-squares policy iteration. IEEE Trans Sustain Energy 2017;8(4):1618–28. 10.
[151] Chun Wei, Zhang Zhe, Qiao Wei, Qu Liyan. Reinforcement-learning-based intelligent maximum power point tracking control for wind energy conversion systems. IEEE Trans Ind Electron 2015;62(10):6360–70. 10.
[152] Zhang Xiaoshun, Li Shengnan, He Tingyi, Yang Bo, Tao Yu, Li Haofei, Jiang Lin, Sun Liming. Memetic reinforcement learning based maximum power point tracking design for pv systems under partial shading condition. Energy 2019;174:1079–90. 5.
[153] Saenz-Aguirre Aitor, Zulueta Ekaitz, Fernandez-Gamiz Unai, Lozano Javier, Lopez-Guede Jose Manuel. Artificial neural network based reinforcement learning for wind turbine yaw control. Energies 2019;12(3):436. 1.
[154] Remani T, Jasmin EA, Imthias Ahamed TP. Residential load scheduling with renewable generation in the smart grid: a reinforcement learning approach. IEEE Syst J 2019;13(3):3283–94. 9.
[155] Gregor P. Henze and Jobst Schoenmann. Evaluation of reinforcement learning control for thermal energy storage systems. HVAC R Res 2003;9(3):259–75. 7.
[156] Vázquez-Canteli José R, Ulyanin Stepan, Kämpf Jérôme, Nagy Zoltán. Fusing tensorflow with building energy simulation for intelligent energy management in smart cities. Sustain Cities and Soc 2019;45(243–257). 2.
[157] Rayati Mohammad, Sheikhi Aras, Ali Mohammad Ranjbar. Optimising operational cost of a smart energy hub, the reinforcement learning approach. Int J Parallel, Emergent Distributed Syst 2015;30(4):325–41. 7.
[158] Mocanu Elena, , Decebal Constantin Mocanu, Nguyen Phuong H, Liotta Antonio, Webber Michael E, Gibescu Madeleine, Slootweg JG. On-line building energy optimization using deep reinforcement learning. IEEE Trans Smart Grid 2019;10(4):3698–708. 7.
[159] Odonkor Philip, Lewis Kemper. Automated design of energy efficient control strategies for building clusters using reinforcement learning. J Mech Des 2019;141(2):2 [Online; accessed 2019-12-11].
[160] Tang Yufei, Yang Jun, Yan Jun, He Haibo. Intelligent load frequency controller using gradp for island smart grid with electric vehicles and renewable resources. Neurocomputing 2015;170:406–16. 12.
[161] Ko Haneul, Pack Sangheon, Victor C, Leung M. Mobility-aware vehicle-to-grid control algorithm in microgrids. IEEE Trans Intell Transport Syst 2018;19(7):2165–74. 7.
[162] Dang Qiyun, Wu Di, Benoit Boulet. A q-learning based charging scheduling scheme for electric vehicles. IEEE transportation electrification conference and expo (ITEC), 6 2019. 2019. p. 1–5. ISSN: 2377-5483.
[163] Levine S, Wagener N, Abbeel P. Learning contact-rich manipulation skills with guided policy search. 2015. arXiv preprint arXiv:1501.05611.
[164] Marc Deisenroth and Carl E Rasmussen Pilco: A model-based and data-efficient approach to policy search. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 465–472, 2011.
[165] Perera ATD, Nik Vahid M, Mauree Dasaraden, Scartezzini Jean-Louis. Electrical hubs: an effective way to integrate non-dispatchable renewable energy sources with minimum impact to the grid. Appl Energy 3 2017;190(232–248).

[166] Perera ATD, Attalage RA, Perera KKCK, Dassanayake VPC. Designing standalone hybrid energy systems minimizing initial investment, life cycle cost and pollutant emission. Energy 2013;54:220–30. 6.
[167] Perera ATD, Attalage RA, Perera KKCK, Dassanayake VPC. A hybrid tool to combine multi-objective optimization and multi-criterion decision making in designing standalone hybrid energy systems. Appl Energy 2013;107:412–25. 7.
[168] Alexander Buttler, Spliethoff Hartmut. Current status of water electrolysis for energy storage, grid balancing and sector coupling via power-to-gas and power-to-liquids: a review. Renew Sustain Energy Rev 2018;82:2440–54. 2.
[169] Anschel Oron, Baram Nir, Nahum Shimkin. Averaged-dqn: variance reduction and stabilization for deep reinforcement learning. International conference on machine learning. PMLR; 2017. p. 176–85.
[170] Papini Matteo, Binaghi Damiano, Canonaco Giuseppe, Pirotta Matteo, Restelli Marcello. Stochastic variance-reduced policy gradient. 2018. arXiv preprint arXiv:1806.05618.
[171] Iyengar Garud N. Robust dynamic programming. Math Oper Res 2005;30(2):257–80.
[172] Daniel J. Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Timothy A. Mann, Todd Hester, and Martin A. Riedmiller. Robust reinforcement learning for continuous control with model misspecification. CoRR, abs/1906.07516, 2019.
[173] Di-Castro Shashua Shirli, Mannor Shie. Deep robust kalman filter. 2017. CoRR, abs/1703.02310.
[174] Chen Tessler, Efroni Yonathan, Mannor Shie. Action robust reinforcement learning and applications in continuous control. 2019.
[175] Pinto Lerrel, Davidson James, Sukthankar Rahul, Gupta Abhinav. Robust adversarial reinforcement learning. International conference on machine learning. 2017. p. 2817–26.
[176] Kamalaruban Parameswaran, Huang Yu-Ting, Hsieh Ya-Ping, Paul Rolland, Shi Cheng, Cevher Volkan. Robust reinforcement learning via adversarial training with Langevin dynamics. 2020. arXiv preprint arXiv:2002.06063.
[177] Jaderberg Max, Mnih Volodymyr, , Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, Silver David, Kavukcuoglu Koray. Reinforcement learning with unsupervised auxiliary tasks. 2016. arXiv preprint arXiv:1611.05397.
[178] Gelada Carles, Kumar Saurabh, Jacob Buckman, Nachum Ofir, Bellemare Marc G. Deepmdp: learning continuous latent space models for representation learning. 2019. arXiv preprint arXiv:1906.02736.
[179] François-Lavet Vincent, Bengio Yoshua, Precup Doina, Pineau Joelle. Combined reinforcement learning via abstract representations. Proceedings of the AAAI conference on artificial intelligence, vol. 33; 2019. p. 3582–9.
[180] Bellemare Marc, Dabney Will, Dadashi Robert, Ali Taiga Adrien, Samuel Castro Pablo, Le Roux Nicolas, Dale Schuurmans, Lattimore Tor, Clare Lyle. A geometric perspective on optimal representations for reinforcement learning. Advances in neural information processing systems. 2019. p. 4358–69.
[181] Kurutach Thanard, Aviv Tamar, Ge Yang, Russell Stuart J, Abbeel Pieter. Learning plannable representations with causal infogan. Advances in neural information processing systems. 2018. p. 8733–44.
[182] Anand Ankesh, Racah Evan, Ozair Sherjil, Bengio Yoshua, Marc-Alexandre Côté, Hjelm R Devon. Unsupervised state representation learning in atari. Advances in neural information processing systems. 2019. p. 8769–82.
[183] Srinivas Aravind, Laskin Michael, Abbeel Pieter. Curl: contrastive unsupervised representations for reinforcement learning. 2020. arXiv preprint arXiv:2004.04136.
[184] Agarwal Alekh, Kakade Sham M, D Lee Jason, Mahajan Gaurav. Optimality and approximation with policy gradient methods in markov decision processes. 2019. arXiv preprint arXiv:1908.00261.
[185] Zhang Kaiqing, Koppel Alec, Zhu Hao, Başar Tamer. Global convergence of policy gradient methods to (almost) locally optimal policies. 2019. arXiv preprint arXiv:1906.08383.
[186] Chen Yichen, Wang Mengdi. Stochastic primal-dual methods and sample complexity of reinforcement learning. 2016. arXiv preprint arXiv:1612.02516.
[187] Lee Donghwan, He Niao. Stochastic primal-dual q-learning. 2018. arXiv preprint arXiv:1810.08298.
[188] Yee Teh, Bapst Victor, Wojciech M Czarnecki, Quan John, Kirkpatrick James, Hadsell Raia, Heess Nicolas, Pascanu Razvan. Distral: robust multitask reinforcement learning. Advances in neural information processing systems. 2017. p. 4496–506.
[189] Espeholt Lasse, Hubert Soyer, Munos Remi, Simonyan Karen, Mnih Volodymir, Ward Tom, Yotam Doron, Vlad Firoiu, Tim Harley, Dunning Iain, et al. Impala: scalable distributed deep-rl with importance weighted actor-learner architectures. 2018. arXiv preprint arXiv:1802.01561.
[190] Parisotto Emilio, Lei Ba Jimmy, Salakhutdinov Ruslan. Actor-mimic: deep multitask and transfer reinforcement learning. 2015. arXiv preprint arXiv:1511.06342.
[191] Calandriello Daniele, Lazaric Alessandro, Restelli Marcello. Sparse multi-task reinforcement learning. Advances in neural information processing systems. 2014. p. 819–27.
[192] Tutunov Rasul, Kim Dongho, Ammar Haitham Bou. Distributed multitask reinforcement learning with quadratic convergence. Advances in neural information processing systems. 2018. p. 8907–16.
[193] Yuan Jianjun, Lamperski Andrew. Online convex optimization for cumulative constraints. Advances in neural information processing systems. 2018. p. 6137–46.
[194] Li Yingying, Qu Guannan, Li Na. Online optimization with predictions and switching costs: fast algorithms and the fundamental limit. 2018. arXiv preprint arXiv:1801.07780.
[195] Lesage-Landry Antoine, Taylor Joshua A. Online convex optimization for demand response. Proc. Bulk Power Syst. Dyn Contr Symp. 2017:1–8.
[196] Cheng Ching-An, Combes Remi Tachet des, Boots Byron, Gordon Geoff. A reduction from reinforcement learning to no-regret online learning. 2019. arXiv preprint arXiv:1911.05873.
[197] Vieillard Nino, Olivier Pietquin, Geist Matthieu. On connections between constrained optimization and reinforcement learning. 2019. arXiv preprint arXiv:1910.08476.
[198] Bu Lucian, Babu Robert, De Schutter Bart, et al. A comprehensive survey of multiagent reinforcement learning. IEEE Trans Syst Man, and Cybernet, Part C (Applications and Reviews) 2008;38(2):156–72.
[199] Hernandez-Leal Pablo, Kartal Bilal, Taylor Matthew E. A survey and critique of multiagent deep reinforcement learning. Aut Agents Multi-Agent Syst 2019;33(6):750–97.
[200] Thanh Thi Nguyen, Nguyen Ngoc Duy, Nahavandi Saeid. Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications. IEEE Transactions on Cybernetics; 2020.
[201] Zhang Kaiqing, Yang Zhuoran, Başar Tamer. Multi-agent reinforcement learning: a selective overview of theories and algorithms. 2019. arXiv preprint arXiv:1911.10635.
[202] Lee Donghwan, He Niao, Kamalaruban Parameswaran, Cevher Volkan. Optimization for reinforcement learning: from single agent to cooperative agents. 2019. arXiv preprint arXiv:1912.00498.
[203] Papoudakis Georgios, Christianos Filippos, Rahman Arrasy, Albrecht Stefano V. Dealing with non-stationarity in multi-agent deep reinforcement learning. 2019. arXiv preprint arXiv:1906.04737.
[204] Tampuu Ardi, Matiisen Tambet, Kodelja Dorian, Kuzovkin Ilya, Korjus Kristjan, Aru Juhan, Aru Jaan, Vicente Raul. Multiagent cooperation and competition with deep reinforcement learning. PloS One 2017;12(4).
[205] Bansal Trapit, Pachocki Jakub, Sidor Szymon, Sutskever Ilya, Mordatch Igor. Emergent complexity via multi-agent competition. 2017. arXiv preprint arXiv:1710.03748.
[206] Littman Michael L. Value-function reinforcement learning in markov games. Cognit Syst Res 2001;2(1):55–66.
[207] Doan Thinh T, , Siva Theja Maguluri, Romberg Justin. Convergence rates of distributed td (0) with linear function approximation for multi-agent reinforcement learning. 2019. arXiv preprint arXiv:1902.07393.
[208] Kar Soummya, José MF Moura, Vincent Poor H. Qd-learning: a collaborative distributed strategy for multi-agent reinforcement learning through consensus. 2012. arXiv preprint arXiv:1205.0047.
[209] Zhang Kaiqing, Yang Zhuoran, Han Liu, Zhang Tong, Başar Tamer. Fully decentralized multi-agent reinforcement learning with networked agents. 2018. arXiv preprint arXiv:1802.08757.
[210] Nedich Angelia, et al. Convergence rate of distributed averaging dynamics and optimization in networks. Found Trend Syst Contr 2015;2(1):1–100.