# Optimal Management of Grid-Interactive Efficient Buildings via Safe Reinforcement Learning

Xiang Huo[†], Boming Liu[‡], Jin Dong[‡], Jianming Lian[‡], and Mingxi Liu[§]



***Abstract*—Reinforcement learning (RL)-based methods have achieved significant success in managing grid-interactive efficient buildings (GEBs). However, RL does not carry intrinsic guarantees of constraint satisfaction, which may lead to severe safety consequences. Besides, in GEB control applications, most existing safe RL approaches rely only on the regularisation parameters in neural networks or penalty of rewards, which often encounter challenges with parameter tuning and lead to catastrophic constraint violations. To provide enforced safety guarantees in controlling GEBs, this paper designs a physics-inspired safe RL method whose decision-making is enhanced through safe interaction with the environment. Different energy resources in GEBs are optimally managed to minimize energy costs and maximize customer comfort. The proposed approach can achieve strict constraint guarantees based on prior knowledge of a set of developed hard steady-state rules. Simulations on the optimal management of GEBs, including heating, ventilation, and air conditioning (HVAC), solar photovoltaics, and energy storage systems, demonstrate the effectiveness of the proposed approach.**

***Index Terms*—Distributed energy resources, grid-interactive efficient buildings, reinforcement learning, safe learning**

## I. Introduction

The optimal management of grid-interactive efficient buildings (GEBs) is becoming more crucial than ever in reducing energy costs, integrating renewables, and facilitating grid decarbonization, owing to the fast-paced deployment of distributed energy resources (DERs). DERs, such as energy storage system (ESS), solar photovoltaic (PV), and electric vehicle (EV), together with flexible loads in GEBs, can be optimally controlled to provide revolutionary improvements in energy efficiency, grid resilience, and cost savings [1], [2]. However, the growing modeling and computing complexity in GEB control, especially the explosion of DERs, is posing unprecedented challenges for the safe and efficient management of GEBs.

To achieve the optimal management of GEBs, model-based methods that leverage scalable architectures have been thoroughly established [3]–[5]. In [3], a decentralized control method was proposed to regulate the voltage in radial distribution networks by coordinating on-load tap changing transformers and PV inverters. Pan *et al.* [4] proposed a distributed low-communication algorithm for the control of islanded series of PV-ESS-hybrid systems. In [5], a two-facet scalable distributed algorithm was developed to offer scalability over both the agent population size and network dimension and verified through a residential EV charging control problem. In [6], an affinely adjustable robust extension of the distributed alternating direction method of multipliers (ADMM) algorithm was designed to compensate for the deviations of forecasted loads and PV generation. Although model-based methods, including both distributed and decentralized control strategies, offer high scalability in GEB control problems, they can encounter impaired model accuracy and increased computing complexity, especially in complex building environments.

To deal with the growing system complexity, reinforcement learning (RL) has been prominently studied in power system applications owing to its model-free nature [7]. RL-based control methods have the capability to comprehend hard-to-model dynamics and surpass model-based methods in managing highly complex GEBs. In [8], a real-time model-free autonomous energy management strategy was proposed for residential multi-energy systems based on deep RL, where the user's energy cost is minimized under system uncertainties. In [9], an actor–critic deep RL algorithm was developed based on reward shaping to manage the residential energy consumption profile with limited knowledge of the uncertain factors. In [10], the control of the HVAC system is framed as a Markov decision process (MDP) using deep neural networks, and solved with deep deterministic policy gradient method to identify the optimal control strategy that reduces the energy cost and improves user comfort.

The aforementioned RL-based approaches learn from trial-and-error interactions with the environment to maximize certain rewards. However, they generically neglect the safety considerations, i.e., ensure the safe behavior of RL agents in their environment. Consequently, the neglect of unsafe actions can lead to adverse consequences throughout both the online training and policy execution. Recently, safe learning methods are gaining more attention in industrial cyber-physical

*This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan . This work has been supported in part by DOE's Office of Electricity, in part by DOE's Building Technologies Office, and in part by NSF Award: ECCS-2145408.

[†]Xiang Huo was with the University of Utah, Salt Lake City, UT 84112 USA. He is now with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: xiang.huo@tamu.edu).

[‡]Boming Liu, Jin Dong, and Jianming Lian are with the Electrification and Energy Infrastructures Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (e-mail: liub,dongj,lianj@ornl.gov).

[§]Mingxi Liu is with the Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT 84112 USA (e-mail: mingxi.liu@utah.edu).

systems to emphasize the needs for real-world applications. A representative safe RL algorithm was proposed in [11] to learn a conservative safety estimate of environment states through a critic, with the probability of failure in safety being upper-bounded. Admittedly, safe RL algorithms balance the tradeoff between safety and policy improvement, however, practical learing-based GEB control that offer customized or strict guarantees on certain hard constraints is still premature.

To fill this research gap, we aim at developing an RL-based GEB control algorithm with enhanced safety guarantees. The contributions of this paper are three-fold: 1) We develop an RL-aided GEB management framework to achieve optimized energy cost and consumer satisfaction; 2) The proposed physics-inspired safe RL minimizes room temperature constraint violations by projecting the actions onto a feasible region formed via a set of steady-state hard constraints; 3) The proposed method exhibits computational efficiency owing to its low computing cost in formulating steady-state feasible regions. Additionally, it is flexible in hard constraint selection and can be embedded in different RL structures.

## II. System Model and Algorithm Design

### A. System Model

*1) Dynamic building thermal network model:* The thermal dynamics are given based on the control-oriented resistance-capacitance (RC) thermal network model for a residential house with an attic [12], [13]. The thermal parameters are identified via a 4R4C thermal network model. The heat transfer within the building can be represented by the following system of first-order differential equations as

$$C_{in}\frac{dT_t^{in}}{dt} = \frac{T_t^w - T_t^{in}}{R_w/2} + \frac{T_t^a - T_t^{in}}{R_a} + \frac{T_t^m - T_t^{in}}{R_m} + \frac{T_t^{amb} - T_t^{in}}{R_{win}} + Q_t^{IHL} - c_1 Q_t^{AC} + c_2 Q_t^{sol} \tag{1a}$$

$$C_w\frac{dT_t^w}{dt} = \frac{T_t^{sol,w} - T_t^w}{R_w/2} - \frac{T_t^w - T_t^{in}}{R_w/2} \tag{1b}$$

$$C_a\frac{dT_t^a}{dt} = \frac{T_t^{sol,f} - T_t^a}{R_f} + \frac{T_t^a - T_t^{in}}{R_a} + c_4 T_t^{sol,a} \tag{1c}$$

$$C_m\frac{dT_t^m}{dt} = -\frac{T_t^m - T_t^{in}}{R_m} + c_3 Q_t^{sol} - c_5 Q_t^{AC} \tag{1d}$$

where $t$ denotes the time index, $C_w$, $C_{in}$, $C_m$, and $C_a$ denote the equivalent overall thermal capacitance of the exterior wall, indoor air, internal mass, and air in attic, respectively. $R_w$, $R_a$, $R_m$, $R_{win}$, and $R_f$ denote the equivalent overall thermal resistance of exterior walls, attic floor, internal mass, window, and roof, respectively. $T_t^{in}$, $T_t^w$, $T_t^a$, $T_t^m$, $T_t^{amb}$, $T_t^{sol,w}$, $T_t^{sol,f}$, and $T_t^{sol,a}$ are the indoor temperature, exterior wall temperature, attic air temperature, internal thermal mass temperature, outdoor dry bulb temperature, and the effects of solar radiation on exterior walls, roofs, and attics, respectively. $c_1$, $c_2$, $c_3$, and $c_4$ denote the effective heating/cooling gain coefficients. $Q_t^{AC}$ denotes the cooling supply of the HVAC, $Q_t^{IHL}$ and $Q_t^{sol}$ denote the sensible heat gain from indoor heat sources and the solar radiation through windows, respectively.

Assume each house is controlled by one HVAC, then the cooling supply of the $j$th HVAC from the $j$th house should stay within

$$0 \le Q_{j,t}^{AC} \le \overline{Q}_{AC}, \forall j \in \Omega_m, \forall t \in \Omega_t \tag{2}$$

where $\overline{Q}_{AC}$ denote the cooling capacity limit, and $\Omega_m$ and $\Omega_t$ represent the set of houses and the set of time slots, respectively.

To guarantee the indoor air temperature stays within the user's comfort range, $T_t^{in}$ should satisfy

$$\underline{T}^{in} \le T_t^{in} \le \overline{T}^{in}, \forall j \in \Omega_m, \forall t \in \Omega_t \tag{3}$$

where $\underline{T}^{in}$ and $\overline{T}^{in}$ denote the lower and upper temperature limits, respectively.

*2) Solar photovoltaic:* Solar PV can provide renewable power supply to an GEB by converting solar power into electricity, the active power injection $p_{j,t}^s$ from the PVs connected at house $j$ is limited by

$$0 \le p_{j,t}^s \le \overline{p}_{j,t}^s, \forall j \in \Omega_m, \forall t \in \Omega_t \tag{4}$$

where $p_{j,t}^s$ denotes the active power injection, $\bar{p}_{j,t}^s$ denotes the maximum available active power from the associated PV inverter, and $\bar{p}_{j,t}^s$ is assumed to be known by the forecast.

*3) Energy storage system:* Assume the $j$th house is connected with an ESS that can offer the backup power supply through controlled charging and discharging actions. Let $p_{j,t}^e$ denote the controllable charging (positive) or discharging (negative) power that is in the range of

$$\underline{p}_j^{dch} \le p_{j,t}^e \le \overline{p}_j^{ch}, \forall j \in \Omega_m, \forall t \in \Omega_t \tag{5}$$

where $\underline{p}_j^{dch} < 0$ and $\overline{p}_j^{ch} > 0$ denote the maximum discharging and charging power limits, respectively.

Besides, the capacity $E_{j,t}$ of the $j$th ESS subjects to the capacity limitation of

$$\underline{E}_j \le E_{j,t} \le \overline{E}_j, \forall j \in \Omega_m, \forall t \in \Omega_t \tag{6}$$

where $\underline{E}_j$ denotes the allowable minimum energy stored in the ESS and $\overline{E}_j$ denotes the maximum energy capacity of the ESS.

*4) Objectives:* The objectives of the GEB control problem are formulated at optimizing the cost and comfort factors by 1) minimizing the total energy cost in the building; and 2) regulating the indoor room temperature to align closely with the predetermined setpoints. In specific, the energy cost minimization objective can be written as

$$\mathcal{C}^{pr} = \sum_{j\in\Omega_m}\sum_{t\in\Omega_t}\left(p_{j,t}^h + p_{j,t}^e - p_{j,t}^s\right)\Delta t T_t^{pr} \tag{7}$$

where $p_{j,t}^h = \delta Q_{j,t}^{AC}$ denotes the power consumption of the HVAC, $\delta$ denotes the cooling coefficient of performance, $\Delta t$ denotes the time interval, and $T_t^{pr}$ denotes the market electricity price.

The customer comfort is measured by the indoor room temperature deviation as

$$\mathcal{C}^{tem} = \sum_{j\in\Omega_m}\sum_{t\in\Omega_t}\left(T^{in}_{j,t} - T^{set}_{j,t}\right)^2 \tag{8}$$

where $T^{set}_{j,t}$ denotes the predetermined temperature setpoint of house (customer) $j$.

### *B. Algorithm Design*

*1) Markov decision process:* To achieve the GEB optimization objectives while satisfying the system constraints, we formulate the energy building control problem as a Markov decision process (MDP) that aims at optimizing the energy cost and customer's dissatisfaction. In specific, the MDP comprises four key components: a collection of states that represent the environment, a range of potential actions for each state, a reward function that evaluates the value of actions taken in specific states, and the transition probability among different states.

*1) State:* The system state denoted by $s_t$ comprises system status that influence the decision-making process. In the GEB control problem, the system state at any time slot $t$ is defined as

$$s_t = {}_{\times j\in\Omega_m}\left(T^{in}_{j,t}, T^{set}_{j,t}, T^{pr}_t, T^{amb}_{j,t}, \overline{p}^s_{j,t}, E_{j,t}\right). \tag{9}$$

where $\times$ denotes the collection of all elements in a set.

*2) Actions:* Actions denoted by $a_t$ are operational decisions for controlling HVAC and DERs, and are defined by

$$a_t = {}_{\times j\in\Omega_m}\left(Q^{AC}_{j,t}, p^s_{j,t}, p^e_{j,t}\right). \tag{10}$$

*3) Reward:* The reward consists of the negative sum of the electricity costs, the room temperature deviations, and the violation of the DER operational constraints

$$r_t = -\alpha_1 C^{pr}_t - \alpha_2 C^{tem}_t - \alpha_3 C^s_t - \alpha_4 C^{cd}_t - \alpha_5 C^{ess}_t \tag{11}$$

where $\alpha_\varpi$, $\varpi = 1,\ldots,5$, denotes the associated penalty coefficients.

The first term $C^{pr}_t$ calculates the energy cost by

$$C^{pr}_t = \sum_{j\in\Omega_m}\left(p^h_{j,t} + p^e_{j,t} - p^s_{j,t}\right)\Delta t T^{pr}_t. \tag{12}$$

The second term $C^{tem}_t$ measures the temperature deviation by

$$C^{tem}_t = \sum_{j\in\Omega_m}\left(T^{in}_{j,t} - T^{set}_{j,t}\right)^2. \tag{13}$$

The third term $C^s_t$ reflects the violation of solar PV limit via

$$C^s_t = \sum_{j\in\Omega_m}\left(p^s_{j,t} - \bar{p}^s_{j,t}\right)^2. \tag{14}$$

The fourth and fifth terms $C^{cd}_t$ and $C^{ess}_t$ assess violations on the ESS charging/discharging power limits and ESS capacity limit, respectively, and they are calculated by

$$\begin{aligned} C^{cd}_t &= \sum_{j\in\Omega_m}\max(0, p^e_{j,t} - \overline{p}^{ch}_j) + \max(0, \underline{p}^{dch}_j - p^e_{j,t}) \\ C^{ess}_t &= \sum_{j\in\Omega_m}\max(0, E_{j,t} - \overline{E}_j) + \max(0, \underline{E}_j - E_{j,t}). \end{aligned} \tag{15}$$

In an attempt to optimize cost-effectiveness, ESSs will strategically charge when market electricity prices are low, while inversely, discharge is preferred during periods of high electricity prices. The solar PVs are stimulated through rewards to inject the solar energy into the grid to meet the load demand. Any superfluous solar energy generation will be stored in the ESS for future utilization. The metric of customer dissatisfaction is quantitatively represented as the discrepancy between the actual room temperature and the scheduled temperature setpoints.

*2) Steady-state analysis:* Inspired by the physics of HVAC system, we propose a safe RL-based learning method that incorporates physics principles to achieve enhanced safety guarantees. The heat transfer within the building is first formulated into a steady-state problem, aiming at tracking desired temperature setpoints. Then we develop a set of indoor room temperature constraints based on the steady-state analysis to regulate the constraint violations. The cooling supply from HVAC will be controlled to drive the system states to a series of steady states that are the optimal solutions. For simplicity, we only consider the summer cooling scenario when the HVACs are always on.

**Proposition 1**. When the HVAC system is on, the linear system in (1) converges asymptotically, and the equilibrium state is uniquely determined by the input $\mathbb{T} \triangleq \{Q^{AC}_t, T^{amb}_t, Q^{IHL}_t, Q^{sol}_t, T^{sol,w}_t, T^{sol,f}_t, T^{sol,a}_t\}$. □

Proposition 1 can be verified by representing the linear system in (1) into a continuous state space model as

$$\dot{\boldsymbol{\mathcal{X}}}_{\tilde{t}} = \boldsymbol{A}\boldsymbol{\mathcal{X}}_{\tilde{t}} + \boldsymbol{B}Q^{AC}_{\tilde{t}} + \boldsymbol{G}\boldsymbol{\mathcal{D}}_{\tilde{t}} \tag{16a}$$

$$\boldsymbol{\mathcal{Y}}_{\tilde{t}} = \boldsymbol{C}\boldsymbol{\mathcal{X}}_{\tilde{t}} \tag{16b}$$

where

$$\boldsymbol{A} = \begin{bmatrix} \frac{-2}{R_w} - \frac{1}{R_a} - \frac{1}{R_m} - \frac{1}{R_{win}} & \frac{2}{R_w} & \frac{1}{R_a} & \frac{1}{R_m} \\ \frac{2}{R_w} & \frac{-4}{R_w} & 0 & 0 \\ \frac{1}{R_a} & 0 & \frac{1}{R_a} - \frac{1}{R_f} & 0 \\ \frac{1}{R_m} & 0 & 0 & \frac{-1}{R_m} \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} -c_1 \\ 0 \\ 0 \\ -c_5 \end{bmatrix}, \boldsymbol{G} = \begin{bmatrix} \frac{1}{R_{win}} & 1 & c_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{2}{R_w} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{R_f} & c_4 \\ 0 & 0 & c_3 & 0 & 0 & 0 \end{bmatrix},$$

$\boldsymbol{\mathcal{X}}_{\tilde{t}} = [T^{in}_{\tilde{t}}, T^w_{\tilde{t}}, T^a_{\tilde{t}}, T^m_{\tilde{t}}]^\intercal$ denotes the system state, $\boldsymbol{\mathcal{Y}}_{\tilde{t}}$ denotes the system output, $\boldsymbol{\mathcal{D}}_{\tilde{t}} = [T^{amb}_{\tilde{t}}, Q^{IHL}_{\tilde{t}}, Q^{sol}_{\tilde{t}}, T^{sol,w}_{\tilde{t}}, T^{sol,f}_{\tilde{t}}, T^{sol,a}_{\tilde{t}}]^\intercal$ denotes the constant environment input, $\boldsymbol{C} \in \mathbb{R}^{4\times 4}$ denotes an identity matrix, and $\tilde{t}$ denotes the steady-state time index for clarity. The controllability of system (16) can be directly verified using Kalman's test.

Therefore, when the HVAC system is on, the objective aims at designing the dynamics of $Q^{AC}_{\tilde{t}}$ to achieve desired steady states that can track the predefined temperature points [14]. The HVAC supply is controlled as input to maintain the indoor air temperature as close as possible to the setpoint. At any steady state, (16a) becomes

$$\boldsymbol{A}\boldsymbol{\mathcal{X}}_{\tilde{t}} + \boldsymbol{B}Q^{AC}_{\tilde{t}} + \boldsymbol{G}\boldsymbol{\mathcal{D}}_{\tilde{t}} = \boldsymbol{0}. \tag{17}$$

Therefore, we can readily obtain that

$$\boldsymbol{\mathcal{X}}_{\tilde{t}} = -\boldsymbol{A}^{-1}(\boldsymbol{B}Q_{\tilde{t}}^{AC} + \boldsymbol{G}\boldsymbol{\mathcal{D}}_{\tilde{t}}). \tag{18}$$

Consequently, the indoor temperature equals to

$$T_{\tilde{t}}^{in} = \frac{c_1 + c_2}{\tau} Q_{\tilde{t}}^{AC} + \frac{\omega}{\tau} \tag{19}$$

where $\tau = 1/(z_2 r_a^2) - 1/r_w - 1/r_a - 1/r_{win}$, $\omega_{\hat{\imath}} = \text{row}_{\hat{\imath}}(G)\mathcal{D}_{\tilde{t}}$, $\forall \hat{\imath} = 1, \ldots, 4$, $\text{row}_{\hat{\imath}}(\cdot)$ denotes the $\hat{\imath}$th row of a matrix, and $\omega = \omega_1 + \omega_2/2 - \omega_3/(r_a z_2) + \omega_4$.

Without loss of generality, we take the room temperature constraint for example to show the regulation on GEB temperature control. The cooling supply of an HVAC system at time $\tilde{t}$ should satisfy

$$\Psi_{\tilde{t}} \triangleq \{Q_{\tilde{t}}^{AC} \mid (19), (2), (3)\} \tag{20}$$

where $\Psi_{\tilde{t}}$ denotes the feasible region of $Q_{\tilde{t}}^{AC}$ at time $\tilde{t}$.

By substituting the cooling capacity limit in (2) and the indoor temperature constraint in (3), the feasible region of the cooling supply $\Psi_{\tilde{t}}$ can be explicitly written into

$$\max\{0, \frac{\tau(\omega + \overline{T}^{in})}{c_1 + c_2}\} \leq Q_{\tilde{t}}^{AC} \leq \min\{\frac{\tau(\omega + \underline{T}^{in})}{c_1 + c_2}, \overline{Q}^{AC}\}. \tag{21}$$

Compared with the original cooling supply constraint in (2), the feasible region in (21) confines the operation of an HVAC to achieve optimal steady states. With the assistance of physical knowledge about the HVAC and the control requirements, the feasible region can be self-defined to regulate the indoor room temperature to the optimal or desired setpoints.

*3) Safe layer design:* In GEB control problems, RL agents often operate in dynamic and uncertain building environments where unsafe actions can lead to detrimental consequences. In this section, an external safety layer is integrated into the RL design to verify the safe actions of HVACs before each execution, ensuring temperature constraint guarantees. We prioritize the safe temperature constraints by using the developed steady-state feasible region to achieve enhanced constraint satisfaction when controlling GEBs.

In specific, the derived steady-state constraints of the cooling supply contour a feasible region, i.e., $\Psi_{\tilde{t}+1}$, of the next action. When the HVAC action stays within $\Psi_{\tilde{t}+1}$, it is considered a safe action. If any action that poses a risk falls outside the feasible region, it will be redirected to the action within the region that has the shortest Euclidean distance. The architecture of the proposed safe RL algorithm is shown in Fig. 1.

Note that the transition to the next state is not observed under the unsafe action, as such an action is not actually executed. However, exploration in such an unfeasible area is allowed. To penalize the potential constraint violations by unsafe actions, the following safe-layer penalty is added as

$$\hat{r}_{\tilde{t}} = -\hat{\alpha}\|\hat{a}_{\tilde{t}} - a_{\tilde{t}}\|_2 \tag{22}$$

where $\hat{a}_{\tilde{t}}$ denotes the safe action in $\Psi_{\tilde{t}+1}$ that has the shortest Euclidean distance to $a_{\tilde{t}}$, i.e., $\min_{\hat{a}\in\Psi_{\tilde{t}+1}} \|\hat{a} - a_{\tilde{t}}\|_2$, and $\hat{\alpha}$

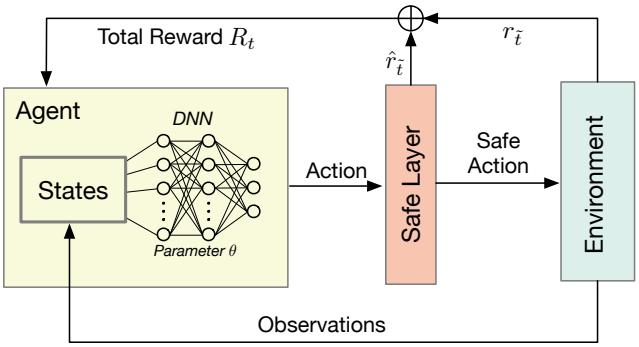


Fig. 1: Physics-inspired safe RL structure.

denotes the penalty coefficient. Note that $\hat{r}_{\tilde{t}}$ expects to train the RL algorithm to prioritize selecting the safe actions for the HVAC, therefore can lead to lower rewards for $r_{\tilde{t}}$.

Therefore, the total reward after adding the safe-layer penalty becomes

$$R_{\tilde{t}} = r_{\tilde{t}} + \hat{r}_{\tilde{t}}. \tag{23}$$

The detailed procedures of the proposed safe RL algorithm are shown in Algorithm 1.

**Remark 1**. The safe layer design of Algorithm 1 is inspired by exploring the thermal dynamics of GEBs in steady states. To apply Algorithm 1, we assume that the hard constraints are prior knowledge that aids in refining the policy and providing safety guarantees. In this paper, the equivalent thermal model RC parameters are obtained through building model identification, and used directly in the environment to demonstrate the performance of safe actions. However, the feasible region $\Psi_{\tilde{t}}$ also can be formulated separately from the RL algorithm based on practical control requirements on the actions. ■

*4) Deep Q-networks:* Q-learning is a model-free RL technique that aims to learn an optimal action-value function, also known as the Q-function, which estimates the expected cumulative reward for taking a particular action in a given state. Deep Q-networks (DQNs) combines deep neural networks with Q-learning by using deep neural networks as function approximators to approximate the Q-function [15].

The state of the system at each time step is represented as inputs to the neural network. In this paper, the state includes room temperatures, outdoor temperatures, electricity price, time of the day, maximum available active power from PVs and the energy statues of the ESSs. The neural network consists of multiple fully connected layers of neurons to take the states as inputs and outputs the Q-values for each possible action. The output layer has one neuron for each possible action, representing the Q-value estimate for each action. During training, it explores the environment by selecting random actions with an exploration rate $\epsilon$. As training progresses, $\epsilon$ decreases and the DQN tends to exploit its learned knowledge by selecting actions with the highest Q-values.

To improve the sample efficiency and reduce correlations between consecutive samples, experience replay is employed. Mini-batches of samples, i.e., $(s_i, a_i, r_i, s_{i+1}, \text{done})$, are ran-

**Algorithm 1:** Physics-inspired GEB control with enhanced safe RL

**Initialization:** Initialize the environment, replay memory $\mathbb{R}$, Q-network, and target Q-network with random weights $\theta$ and $\hat{\theta}$, exploration rate $\epsilon$; Initialize the states: $s_0 = \{T^{in}_{j,0}, T^{set}_{j,0}, T^{pr}_0, T^{amb}_{j,0}, \overline{p}^{s}_{j,0}, E_{j,0}\}$;
**for** $episode = 1, M$ **do**
  Reset the environment $s_{\tilde{t}} = s_0$;
  **for** $\tilde{t} = 1, T$ **do**
    With probability $\epsilon$, select a random action $a_{\tilde{t}}$, otherwise, select $a_{\tilde{t}} = \arg\max_a Q(s_{\tilde{t}}, a; \theta)$;
    **if** $a_{\tilde{t}} \notin \Psi_{\tilde{t}}$ **then**
      $\hat{a}_{\tilde{t}} = \Pi_{\Psi_{\tilde{t}}}(a_{\tilde{t}})$, set $\hat{r}_{\tilde{t}} = -\hat{\alpha}\|\hat{a}_{\tilde{t}} - a_{\tilde{t}}\|_2$;
    **else**
      keep $\hat{a}_{\tilde{t}} = a_{\tilde{t}}$ as the safe action, set $\hat{r}_{\tilde{t}} = 0$;
    **end**
    Execute $\hat{a}_{\tilde{t}}$ on HVAC and obtain the reward $r(s_{\tilde{t}}, \hat{a}_{\tilde{t}})$ and next state $s_{\tilde{t}+1}$ ;
    Store the transition $(s_{\tilde{t}}, a_{\tilde{t}}, r_{\tilde{t}}, s_{\tilde{t}+1}, \text{done})$ in $\mathbb{R}$;
    Sample a mini-batch transitions $(s_i, a_i, r_i, s_{i+1}, \text{done})$ from $\mathbb{R}$ ;
    Set $y_i =$
    $$\begin{cases} r_i & \text{done=1,} \\ r_i + \gamma \max_{a_{i+1}} \hat{Q}(s_{i+1}, a_{i+1}; \hat{\theta}) & \text{done=0;} \end{cases}$$
    Perform gradient descent on $(y_i - Q(s_i, a_i; \theta))^2$;
    Decay $\epsilon$;
    Set $\hat{Q} = Q$ every $k$ steps;
  **end**
**end**

domly sampled from the replay memory $\mathbb{R}$ to train the neural network. By iteratively updating the weights $\theta$ of the neural network based on the training, the model learns to approximate the optimal Q-function and make informed decisions about the HVAC cooling supply, solar PV power injection, and ESS charging/discharging power. The structure of the proposed safe-RL network using DQN is shown in Fig. 2.

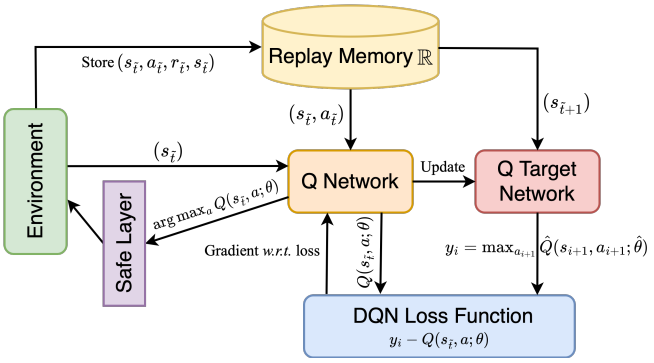


Fig. 2: The structure of the proposed safe-RL network based on DQN.

Algorithm 1 adopts a DQN as the function approximator to calculate the Q-values, the Q-function gives the expected return of taking a particular action $a$ in a given state $s_{\tilde{t}}$. A discount factor $\gamma$ is used to balance the importance of immediate rewards versus future rewards in the decision-making process. With an experienced memory, transitions are stored and sampled randomly for training. During each episode, the agent selects actions based on an $\epsilon$-greedy exploration strategy to balance exploration and exploitation. The Q-network is updated by minimizing the mean squared error loss between the predicted Q-values and the target Q-values, which is $(y_i - Q(s_i, a_i; \theta))^2$. The target network $\hat{Q}$ is updated every $k$ steps by copying the weights from the online network $Q$ to enhance learning stability and improve the convergence of the algorithm. The algorithm iteratively repeats episodes until convergence.

## III. Simulations

In this section, we evaluate the performance of the proposed safe-RL method on an GEB control problem. The objective is to minimize the total energy cost and maximize user comfort as described in (7) and (8). In the simulation setting, one ESS is assumed to be connected at each house with a capacity of 2 kWh and a maximum charging/discharging power of $\pm 1$ kW; Additionally, it is presumed that the ESS has a charging and discharging efficiency of 0.98 and 0.85, respectively, and should preserve allowable minimum energy of 0.3 kWh as backup power for emergency. Solar PV generation was estimated on a sunny day and scaled to have a peak power of 0.3 kW [16].

The time interval is set to be $\Delta t = 15$ mins on a daily basis. The day-ahead electricity price is estimated based on data from the Atlanta electricity market [17]. We adopt the 4R4C model whose RC parameters are identified to be $C_w = 10,000,000$, $C_{in} = 329,472$, $C_m = 14,644,976$, $C_a = 2,330,670$, $R_w = 0.0057$, $R_a = 0.2$, $R_m = 0.1$, $R_{win} = 0.0807$, and $R_f = 0.0965$, respectively. The effective heating/cooling gain coefficients are $c_1 = 0.5$, $c_2 = 0.5$, $c_3 = 0.4$, $c_4 = 0.8$, and $c_5 = 0.5$, respectively. For the training of DQN, the learning rate $\alpha$ is set to be 0.001. The exploration rate is initially set as $\epsilon = 1$ and then linearly decays over time. The discount factor is set as $\gamma = 0.99$. The replay memory size is $10,000$, and the sample batch size for each training iteration is 64. The Q-networks has two hidden layers with 256 hidden units for each layer. Fig. 3 demonstrates the indoor air temperature control results with proposed safe RL algorithm. While the HVAC is on cooling status throughout the simulation, the indoor room temperature is strictly regulated within $[18, 22]$ °C by using the feasible region developed in steady states. Note that the precise temperature control is only obtained as a result of the exact RC model parameters. As shown in Fig. 4, the original actions before entering the safety layer violate the safety bounds drastically. After applying the proposed safe RL, unsafe actions were regulated to stay within the feasible region, leading to enhanced temperature bounds guarantees.

In Fig. 5, the solar power has a peak power generation around noon, and was maximumly utilized in building electricity supply to help reduce the energy cost. Fig. 6 reflects

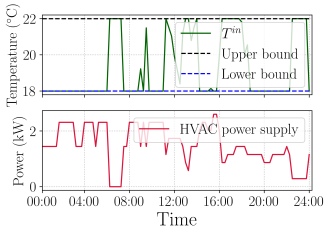


Fig. 3: Indoor room temperature and HVAC power supply.

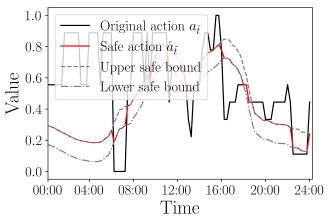


Fig. 4: Original actions and regulated safe actions during the HVAC control.

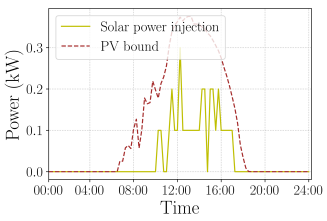


Fig. 5: Solar power injection from the solar PVs.

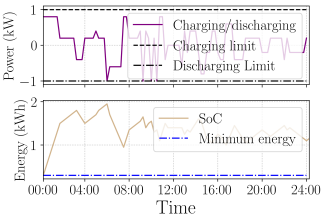


Fig. 6: Charging/discharging schedules of the ESS.

the charging/discharging behaviors of the ESS. The ESS tends to store more energy off the peak hour, e.g., before 8 A.M., and discharge to provide electricity whenever needed. Besides, the capacity of the ESS is always above the lower capacity limit 0.3 kWh. Finally, Fig. 7 measures the performance of

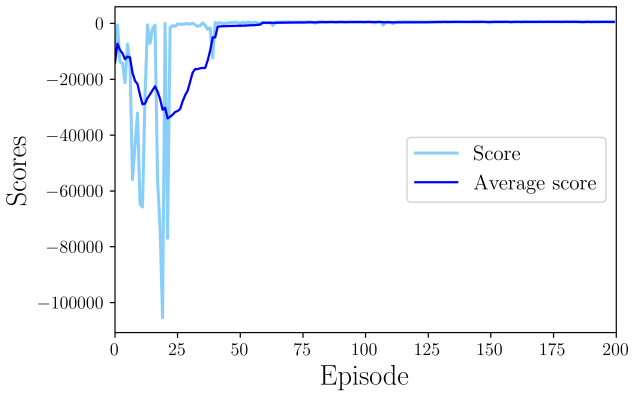


Fig. 7: Score and average scores in the RL process.

the agent's interactions with the environment. The agent learns the optimal GEB control policy that maximizes the cumulative reward over time, including the cumulative score during a specific episode and average scores from all episodes.

## IV. Conclusion

This paper presents a physics-inspired safe-RL framework for the optimal management of DERs and HVACs in GEBs. Different energy-consuming and producing resources are coordinated to optimize energy usage, reduce electricity costs, and improve customer comfort. The proposed safe-RL approach can achieve an enhanced safety constraint guarantee based on the physics of the HVAC system for indoor room temperature control. The developed method also eliminates the computational power overhead compared to solving a dynamic optimization problem. The simulation studies on the GEB problem show the effectiveness of the proposed method in learning to control diverse energy resources in a safe and cost-efficient manner. Future work includes contemplating the trade-off between reward and safety, when applied to large-scale cyber-physical energy systems.

## References


[1] J. R. Aguero, E. Takayesu, D. Novosel, and R. Masiello, "Modernizing the grid: Challenges and opportunities for a sustainable future," *IEEE Power and Energy Magazine*, vol. 15, no. 3, pp. 74–83, 2017.

[2] J. Su, R. Zhang, P. Dehghanian, M. H. Kapourchali, S. Choi, and Z. Ding, "Renewable-dominated mobility-as-a-service framework for resilience delivery in hydrogen-accommodated microgrids," *International Journal of Electrical Power & Energy Systems*, vol. 159, p. 110047, 2024.

[3] M. R. Jafari, M. Parniani, and M. H. Ravanji, "Decentralized control of OLTC and PV inverters for voltage regulation in radial distribution networks with high PV penetration," *IEEE Transactions on Power Delivery*, vol. 37, no. 6, pp. 4827–4837, 2022.

[4] Y. Pan, A. Sangwongwanich, Y. Yang, and F. Blaabjerg, "Distributed control of islanded series PV-battery-hybrid systems with low communication burden," *IEEE Transactions on Power Electronics*, vol. 36, no. 9, pp. 10 199–10 213, 2021.

[5] X. Huo and M. Liu, "Two-facet scalable cooperative optimization of multi-agent systems in the networked environment," *IEEE Transactions on Control Systems Technology*, vol. 30, no. 6, pp. 2317–2332, 2022.

[6] A. Attarha, P. Scott, and S. Thiébaux, "Affinely adjustable robust ADMM for residential DER coordination in distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1620–1629, 2019.

[7] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for selective key applications in power systems: Recent advances and future challenges," *IEEE Transactions on Smart Grid*, 2022.

[8] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3068–3082, 2020.

[9] R. Lu, Z. Jiang, H. Wu, Y. Ding, D. Wang, and H.-T. Zhang, "Reward shaping-based actor-critic deep reinforcement learning for residential energy management," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 3, pp. 2662–2673, 2022.

[10] B. Liu, M. Akcakaya, and T. E. McDermott, "Automated control of transactive HVAC in energy distribution systems," *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2462–2471, 2020.

[11] H. Bharadhwaj, A. Kumar, N. Rhinehart, S. Levine, F. Shkurti, and A. Garg, "Conservative safety critics for exploration," in *International Conference on Learning Representations*, Vienna, Austria, May 4 2021.

[12] B. Cui, J. Joe, J. Munk, J. Sun, and T. Kuruganti, "Load flexibility analysis of residential HVAC and water heating and commercial refrigeration," Oak Ridge National Laboratory, Oak Ridge, TN, United States, Tech. Rep., 2019.

[13] X. Huo, J. Dong, B. Cui, B. Liu, J. Lian, and M. Liu, "Two-level decentralized-centralized control of distributed energy resources in grid-interactive efficient buildings," *IEEE Control Systems Letters*, vol. 7, pp. 997–1002, 2022.

[14] X. Zhang, W. Shi, B. Yan, A. Malkawi, and N. Li, "Decentralized and distributed temperature control via HVAC systems in energy efficient buildings," *arXiv preprint arXiv:1702.03308*, 2017.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[16] National Renewable Energy Laboratory, "NSRDB: National solar radiation database," 2022, https://nsrdb.nrel.gov.

[17] U.S. Energy Information Administration, "Georgia profile state profile and energy estimates," 2023, https://www.eia.gov/state/analysis.php?sid=GA.