Constrained Reinforcement Learning for Building Demand Response

Jerson Sanchez and Jie Cai

Abstract— **This paper presents a constrained reinforcement learning-based control strategy for building demand response. Compared to conventional (unconstrained) reinforcement learning (RL) controllers where indoor comfort constraints are addressed by adding a comfort violation penalty in the reward function, the proposed strategy handles the constraints explicitly, by upper bounding the expected cumulative constraint violation, to avoid the use of arbitrarily set penalty factors that can significantly affect control performance. To demonstrate its efficacy, simulation tests of the proposed strategy as well as baseline model predictive controllers (MPC) and conventional (unconstrained) policy optimization methods were conducted. The simulation tests show that the constrained RL strategy achieved utility cost savings up to 22%, similar to the MPC baselines, with minimum constraint violation, while the unconstrained RL controllers led to either high utility costs or constraint violations, depending on the penalty factor setting.**

I. INTRODUCTION

Building electrical loads represent 75% of the electricity consumption in the United States. Technologies and solutions for efficient and sustainable building operations have witnessed growing attention in the past few years, with heating, ventilation and air conditioning (HVAC) demand response being one of them. Demand response plays an important role in improving electric power system reliability against uncertainties in renewable generation, peak demand, asset availability, and other grid contingent conditions [1]. Model predictive control (MPC) is a broadly used technique in demand response control of HVAC systems. While MPC offers a powerful framework that can explicitly consider constraints during control decision making, a decent process model is needed to achieve satisfactory control performance, the development of which would require significant engineering costs, especially for complex systems such as buildings. Reinforcement learning (RL) offers a promising model-free control approach that has been successfully applied in different fields including building energy management.

1) Related work: Numerous studies have applied RL techniques for building controls. The earliest works used Q-learning techniques and highlighted the challenges of RL techniques that need to be overcome for their practical applications [2]. With the development of new algorithms, such as deep Q-networks (DQN) and policy optimization, better performance could be achieved even for applications in the control of complex systems such as HVAC demand response. For instance, DQN was used with an action processor to leverage previously known information from rule-based controllers to reduce training time [3]. Better performances were shown using deep deterministic policy gradient under a locational marginal price in a multi-zone residential HVAC system [4] and smart home energy management system [5]. DQN strategies have been implemented and demonstrated in real buildings, e.g., a DQN strategy pre-trained by a simulation model was deployed in a residential building demonstrating a potential of 10 to 20% cost savings [6]. Asynchronous advantage actor-critic methods were applied for demand response subject to demand charges in a multi-zone commercial building [7]. Policy optimization algorithms have also been studied in the context of demand response. A two-stage RL training framework was proposed integrating evolutionary strategies and proximal policy optimization (PPO) to address a grid-interactive building control problem [8]. Pure PPO algorithms were also tested under different demand response scenarios in a building simulation tool (EnergyPlus) showing energy reduction of up to 22% and peak demand reduction of up to 50% [9]. All the studies above used unconstrained RL that cannot directly handle constraints, e.g., capacity and indoor comfort constraints, during control decision making. In unconstrained RL implementations, constraints are often addressed heuristically, e.g., by adding a penalty for constraint violations in the reward function. However, the control performance is sensitive to the arbitrarily set penalty factor for indoor discomfort. It is difficult to put a price tag on thermal discomfort as individuals would have different perceptions and tolerances thereof. The local utility rate structures could further complicate the situation as the economic consequence can highly depend on the energy rates as well as the peak-to-off-peak ratio.

2) Contributions of this work: This paper reports the challenges of applying unconstrained RL algorithms to building HVAC control: (1) sensitivity to the choice of the penalty factor, and (2) challenges in benchmarking these algorithms with other techniques such as MPC-based strategies without guaranteed comfort constraint satisfaction. To overcome the challenges, this paper presents a constrained RL-based strategy for building demand response and its numerical test results in comparison with baseline unconstrained RL strategies as well as MPC. To the authors' knowledge, this is the first work that applied constrained RL strategies to tackle these challenges in building demand response.

This paper is structured as follows. First, we introduce the general reinforcement learning concept along with the state-of-the-art algorithms for both unconstrained and constrained policy optimization. Next, we present the numerical building model used as the emulator that interacts with each RL

This material is based upon work supported by the National Science Foundation under Grant No. 2238381 and the Department of Energy under Grant No. DE-SC0023498. The authors are with the School of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, OK, 73072 (jcai@ou.edu).

agent in Section III, followed by brief discussions of the control implementations in Section IV. Section V reports the simulation test results in comparison with the conventional RL and MPC baselines. Concluding remarks are provided in Section VI.

II. Reinforcement Learning

This section introduces the RL framework and algorithms used in this study and formulates the optimization problem that each algorithm seeks to solve.

A. Markov Decision Process

In an ideal case, the states of a dynamic system are all measurable during interactions with the environment where the control problem can be formulated as Markov decision process (MDP). In reality, however, not all states are available to the agent, e.g., wall internal temperatures which characterize building thermal dynamics, and only one or a few of the states are observable, e.g., zone temperatures. Control of such systems can be addressed as a partially observable Markov decision process (POMDP) [10] or as a regular MDP by including historical measurements of the observable outputs and inputs in the state vector. The latter is analogous to the state-space realization of a high-order linear system using the high-order time derivatives of the inputs and outputs as the state variables. This MDP formulation has been used for RL implementations for building HVAC controls [3].

B. Trust Region Policy Optimization (TRPO)

The goal of RL is to obtain a policy π that maximizes the finite or infinite horizon discounted total return $J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$, where τ represents a trajectory and π_θ is a parameterized policy which accepts state observations as inputs and outputs the action probabilities. TRPO [11] seeks to maximize a surrogate objective function subject to a constraint that limits the size of a policy update during each iteration measured by the KL divergence:

$$\max \quad \mathbb{E}_{\tau\sim\pi_{\theta_{old}}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}\hat{A}^{\pi_{\theta_{old}}}(s,a)\right] \tag{1}$$

$$\text{s.t.} \quad \mathbb{E}_{\tau\sim\pi_\theta}\left[D_{KL}(\pi_\theta(\cdot|s)||\pi_{\theta_{old}}(\cdot|s))\right] \le \delta \tag{2}$$

where $\hat{A}^{\pi_\theta}$ is the advantage that can be calculated using the generalized advantage estimation (GAE) method [12] and δ is the upper bound imposed on the KL divergence between two consecutive policy updates. This problem is usually simplified using a linear approximation of the objective function and a quadratic approximation of the KL divergence constraint. Expectations are estimated by a sample mean over trajectories and timesteps in this study. If the agent collects N trajectories of T steps each, the expected value of the objective function in (1) can be estimated as:

$$\frac{1}{NT}\sum_{N}\sum_{t=0}^{T-1}\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}^{\pi_{\theta_{old}}}(s_t,a_t) \tag{3}$$

C. Proximal Policy Optimization (PPO)

PPO seeks to solve the same problem presented in the TRPO case but in a simpler manner. Two versions of PPO algorithms were proposed in the original work [13], while PPO-clip is more broadly used. PPO-clip uses a clip function to limit the incentive of policy change to stabilize training as follows:

$$\max \quad \mathbb{E}_{\tau\sim\pi_{\theta_{old}}} L(s,a,\theta_{old,\theta}) \tag{4}$$

where

$$L(s,a,\theta_{old},\theta) = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}\hat{A}^{\pi_{\theta_{old}}}(s,a),\; h\Big(\epsilon,\hat{A}^{\pi_{\theta_{old}}}(s,a)\Big)\right) \tag{5}$$

and ϵ is a hyper-parameter that determines how far the new policy candidate is allowed to deviate from the old one and the function h is defined as:

$$h\Big(\epsilon,\hat{A}^{\pi_{\theta_{old}}}\Big) = \begin{cases} (1+\epsilon)\hat{A}^{\pi_{\theta_{old}}}, & \hat{A}^{\pi_{\theta_{old}}} > 0 \\ (1-\epsilon)\hat{A}^{\pi_{\theta_{old}}}, & \hat{A}^{\pi_{\theta_{old}}} < 0. \end{cases} \tag{6}$$

Note that both TRPO and PPO can use a parameterized network to approximate the value function $V_\beta(s)$ and compute advantages at each timestep.

D. Constrained Reinforcement Learning

An inherent challenge of non-constrained RL algorithms (e.g., TPRO and PPO) is the inability to address constraints, while for most control applications, operational constraints exist to ensure safe and quality services. A remedy is to incorporate a penalty for possible constraint violations in the reward function but it is difficult to find an appropriate weighting factor between the original merit and constraint satisfaction. While setting a high penalty factor helps better enforce constraints, it limits the improvement in the original objective. On the other hand, setting a low penalty factor prioritizes maximization of the original merit at the expense of significant constraint violations.

Constrained RL seeks to maximize the original merit function but restricts the set of feasible policies so that a discounted constraint violation return $J_C(\pi) = \mathbb{E}_{\tau\sim\pi_\theta}[R_C(\tau)]$ is limited by a set upper bound d [14]. Among the proposed algorithms in the literature, constrained policy optimization (CPO) [15] has been widely studied, which seeks to solve the optimization problem:

$$\max \quad \mathbb{E}_{\tau\sim\pi_{\theta_{old}}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}\hat{A}^{\pi_{\theta_{old}}}(s,a)\right] \tag{7}$$

$$\text{s.t.} \quad J_C(\pi_{\theta_{old}}) + \frac{1}{1-\gamma}\mathbb{E}_{\tau\sim\pi_{\theta_{old}}}\left[\hat{A}_C^{\pi_{\theta_{old}}}(s,a)\right] \le d \tag{8}$$

$$\mathbb{E}_{\tau\sim\pi_\theta}\left[D_{KL}(\pi_\theta||\pi_{\theta_{old}})\right] \le \delta \tag{9}$$

where J_C is the constraint return. The constraint advantage and return are defined in a similar manner to those of the reward function. A positive slack variable d is introduced to stabilize training. CPO uses linear approximations on the objective and the first constraint and a second-order approximation on the KL divergence constraint to solve the optimization problem. CPO uses two parameterized value networks, for the constraint and reward value functions, respectively.

III. Building System Model

This section introduces the building dynamic model used for control testing and MPC synthesis. Note that in the MPC implementation, the same model is assumed for the plant and control synthesis and perfect weather forecast is assumed; therefore, the presented control performance represents the theoretical upper bound and mainly serves the benchmarking purpose.

A discrete-time state-space model is used to reproduce building thermal dynamics, based on a thermal network approach [16]. The model uses the cooling rate as the control input and outputs the zone temperature, in the following form:

$$x^{t+1} = \mathbf{A}x^t + \mathbf{B_w}w^t + \mathbf{B_u}Q_z^t \tag{10}$$

$$T_z^{t+1} = \mathbf{C}x^{t+1} \tag{11}$$

where x^t is the state vector containing all nodal temperatures of a thermal network, Q_z^t is the average cooling rate within each time step, T_z^t is the zone temperature, and w^t is the disturbance vector comprised of outdoor temperature, internal heat gains, and solar radiation. The state-space matrices $\mathbf{A}$, $\mathbf{B_w}$, $\mathbf{B_u}$ and $\mathbf{C}$ are dependent on the thermal resistances and capacitances of the thermal network. Details of the modeling approach and parameter values can be found in [16]. The model shown in (10) and (11) is used as a simulation test bed to evaluate the various control strategies. The same model is adopted for the MPC implementation, while in RL-based control, the system dynamics are not known to the agent and the control policy is learned from the recorded interactions between the agent and the plant model that include control commands sent to the building and the measurable observations. Details of the RL implementation are discussed in Sec. IV-B. The power used by the HVAC system can be estimated from the cooling rate by

$$P^t = \frac{Q_z^t}{COP} \tag{12}$$

where the coefficient of performance (COP) assumes a constant in this study. Equations (10) to (12) define the environment which the agent interacts with. The agent applies a temperature setpoint and a set of measurable observations are available to the agent after execuation of the setpoint. In this study, only a cooling scenario is considered although the methods can be directly applied for heating seasons.

IV. Control Implementations

Five control strategies are considered in this study, i.e., two MPC baselines that represent the current practices for HVAC operations, the TRPO- and PPO-based RL strategies that have been extensively studied in the literature for building control, and the constrained RL strategy.

A. Baseline MPC Control Strategies

Two MPC strategies are formulated as linear programs (LP), with different cost functions and prediction time horizons.

1) Energy Minimization Control (E-Min): The first baseline strategy represents the current practice for maximum energy efficiency. This strategy tends to maintain the zone temperature as close to the upper limit of the comfort temperature zone as possible to minimize the HVAC energy consumption of each time step, with the following cost function:

$$W_1 = P^t. \tag{13}$$

The energy minimization strategy represents a greedy control policy over the HVAC energy use with only one step ahead prediction. A number of operational constraints should be respected in control decision making. Upper and lower limits are imposed for the zone temperature in order to meet the indoor comfort requirements:

$$T_{min}^t \le T_z^t \le T_{max}^t \tag{14}$$

where T_{min} and T_{max} are the lower and upper bounds of the comfort band. These temperature limits can vary with the occupancy status: when the building is occupied, a tighter temperature constraint is needed to ensure comfort, while during unoccupied hours, relaxed temperature bounds can be used to achieve energy or cost savings. The cooling rate at each time step should be bounded by the cooling capacity Q_T, which is assumed to be time-invariant in this study:

$$0 \le Q_z^t \le Q_T. \tag{15}$$

2) Utility Cost Minimization Control (U-Min): The second baseline strategy minimizes the electric utility cost under time-of-use rates over a look-ahead time horizon (e.g., 12 hours in the case study) and represents the current practice for predictive demand responsive control. This strategy can be implemented by solving an LP with a cost function in the following form:

$$W_2 = \sum_{t=t_i}^{t_f} (r^t \cdot P^t) \tag{16}$$

where r^t is the retail energy rate (\$/kWh) that may change with time of the day, t_i and t_f are the first and last time steps of the prediction horizon. The same constraints discussed for the E-min strategy are also present in this cost-minimizing strategy, over the whole look-ahead time horizon.

B. Reinforcement Learning Formulation

The building demand response problem can be formulated as a MPD described in section II-A. The basic elements of the MDP are discussed as follows:

1) State observations: At every 15-min time step, a set of observations is available to the agent, which includes the current zone temperature T_z^t, the past five-step zone temperatures $T_z^{t-1},...,T^{t-5}$, hour of the day h_t, and the 12-hour forecasts of the outdoor temperature $T_{out}^t,...,T_{out}^{t+47}$ and global horizontal solar radiation $q_{sol}^t,...,q_{sol}^{t+47}$. The observation set also contains the power P^t required to achieve the temperature setpoint and the upper and lower zone temperature limits at the current time step.

2) Actions: The action taken by the agent is the zone temperature setpoint for the next decision step.

3) Rewards: The goal of the RL agent is to maximize the reward (negative of the cost) under a time-of-use (TOU) tariff while maintaining the zone temperature within the comfort bounds.

Unconstrained RL: Conventional (unconstrained) RL techniques cannot explicitly handle control constraints. In almost all previous building control applications, the comfort constraints were addressed as a soft penalty cost added to the energy cost with a prescribed weighting factor. The reward associated with the energy cost is $R_{cost} = -r_t P^t$. The reward associated with the comfort violation penalty is considered in the following form:

$$R_{com} = -\phi\big(\max(T_{z,min}^t - T_z^t, 0) + \max(T_z^t - T_{z,max}^t, 0)\big) \quad (17)$$

where ϕ is the constraint violation penalty factor. This comfort penalty is proportional to the cumulative temperature excursions out of the comfort band. The overall reward for the unconstrained RL case is $R_{unc} = R_{cost} + R_{com}$. Note that this penalty does not take into account the occupant's comfort perception and is only based on violations of the set temperature bounds.

Constrained RL: In the constrained RL case the reward and constraint (J_C in Eq (9)) functions assume R_{cost} and $-R_{com}/\phi$, respectively.

V. Case Study Results

Simulation tests were conducted to assess the performance of the constrained RL algorithm in comparison with the different baseline strategies. The TOU retail energy rates used in the simulation were obtained from El Paso Electric Co. [17] and are shown in Table I. A lower energy rate of \$0.07/kWh is involved during non-peak hours while electricity is charged at a much higher rate of \$0.22/kWh during on-peak hours. The zone temperature bounds change with the occupancy of the building, with 9AM to 6PM being the occupied period. During occupied hours the upper and lower temperature bounds are 21.5°C and 23.5°C, while during unoccupied hours are 20.5°C and 24.5°C, respectively.

TABLE I
Summer Time of Use tariff

Electricity price ($/kWh)	Hours
0.222	12:00 to 18:00
0.077	Rest of day

The MPC baselines used a look-ahead horizon of 12 hours with a decision implemented every 15-min time step. The MPC baselines were formulated using the CVX package in MATLAB [18] and solved using Gurobi [19]. The RL strategies were trained using two years of simulation data following an on-policy scheme. The RL baseline simulations were obtained using the Stable Baselines 3 OpenAI library [20] and the CPO algorithm was implemented using PyTorch. Episodic training was utilized with each episode or trajectory consisting of 2 days (192 steps). TRPO and PPO with three different constraint violation penalty factors, i.e. , $\phi = 0.01$, $\phi = 0.1$, and $\phi = 1$, were chosen as RL baselines to illustrate the effect of the weighting factor on the control performance.

Policy networks use the observation set as input, 2 hidden layers of 64 neurons and an output layer with 17 possible action logits. A final softmax layer is used to predict the probability of choosing an action given an observation set. The value networks used by all the RL algorithms and the constraint network involved in the CPO algorithm share the same structure with 2 hidden layers of 64 neurons for each layer and a one-dimensional output layer.

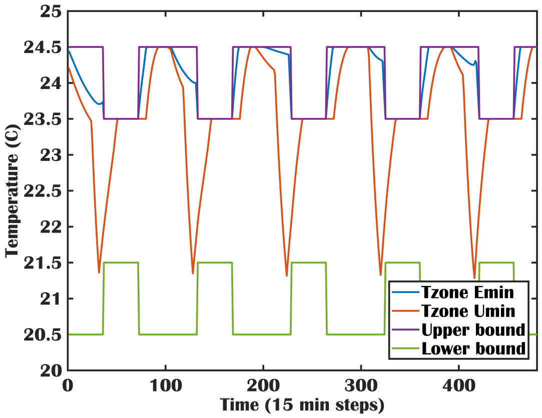

Fig. 1. Zone temperature under MPC baselines.

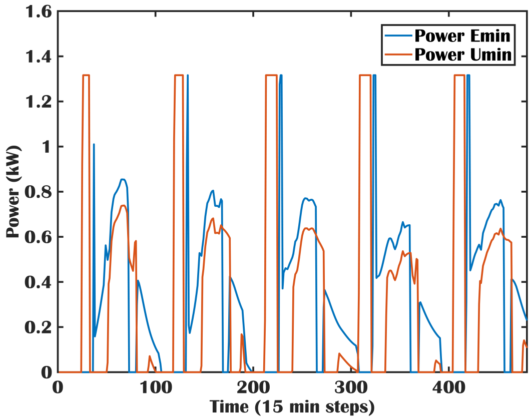

Fig. 2. Power profile under MPC baselines.

Table II shows the energy costs and total energy usage for the five evaluated strategies. Fig. 1 and Fig. 2 present the simulation results of the two MPC baseline strategies. The energy minimization baseline (E-min) maintains the zone temperature at the upper bound when mechanical cooling is called for, resulting in minimum energy usage. During unoccupied hours the temperature floats and the HVAC system remains off. The utility-cost-minimizing (U-min) strategy engages a pre-cooling action before each on-peak period. The pre-cooling action maintains a lower zone temperature prior to on-peak hours so that "cooling" energy is stored in the building thermal mass; during on-peak hours, the zone temperature is adjusted upwards to allow the stored cooling energy to be released, resulting in shifting of building electricity use to low-cost hours to reduce utility cost. While this strategy presents utility cost savings, it increases the total energy used by the HVAC system. Compared to strategy E-min, the cost-minimizing MPC strategy achieves cost savings of 20%, with a total energy rebound of 10%. This represents the best economic performance that can possibly be obtained under the same prediction horizon setting without any comfort constraint violation. Actual MPC performance would be worse due to potential control-plant model mismatches.

Fig. 3 and 4 present the simulation test results under the two unconstrained RL strategies subject to different constraint violation penalty factors. As expected, the unconstrained RL strategy with the lowest comfort penalty factors leads to the lowest energy cost, with cost savings up to 36% relative to E-min, but at the expense of significant comfort issues. The RL strategies with high comfort penalty factors are able to regulate the zone temperature within the comfort zone but lead to high HVAC energy costs (savings of only less than 8% relative to E-min). The high-comfort-penalty RL strategies do not execute any pre-cooling actions for load shifting, which is the major cause of the high energy cost as can be seen in Table II. It is evident from these results that performance achievable with the unconstrained RL strategies is highly dependent on the comfort penalty factor setting. This also makes the comparison with the MPC benchmarks challenging.

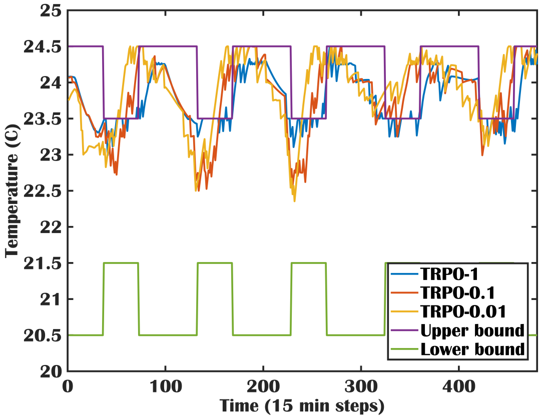

Fig. 3. Zone temperatures under TRPO strategies.

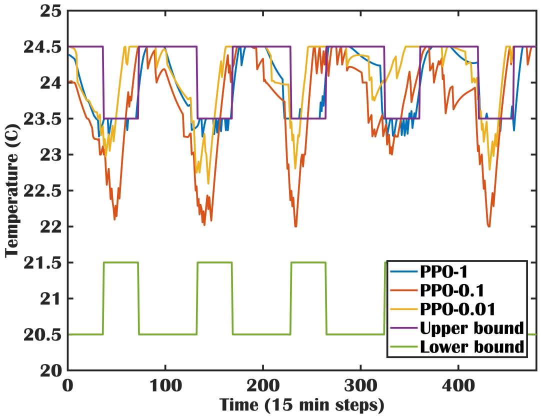

Fig. 4. Zone temperatures under PPO strategies.

Fig. 5 shows the learning curve for the constrained RL algorithm, while the 60-episode average constraint function value is shown in Fig. 6. It can be seen that the constraint violation approaches the set limit $d = 0.1$ and succeeds at enforcing the constraint satisfaction within the set limit. This effect is evident in Fig. 7 which shows the zone temperature variation under the CPO strategy. Pre-cooling actions similar to those in the benchmark MPC strategies are present in the CPO case. The constrained RL strategy achieves utility cost savings of 22.7%, which outperforms the utility-minimizing strategy due to the minor temperature constraint violations. Note that choosing a slack term (d) too close to zero could cause training stability issues and a small positive value is needed to ensure stable and reliable training. These results clearly demonstrate the superior performance of constrained RL over the unconstrained counterparts for predictive control of building flexible loads while ensuring indoor comfort.

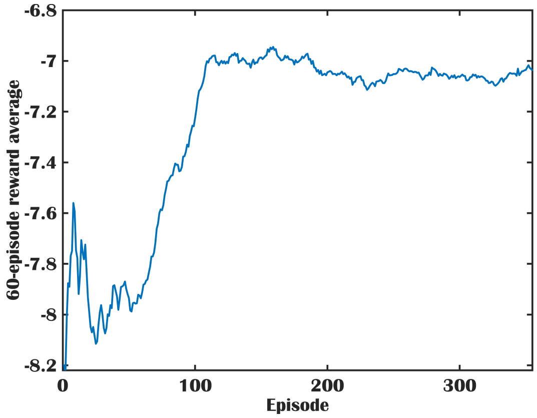

Fig. 5. Moving average reward in CPO training.

VI. Conclusions

This paper presented a constrained reinforcement learning-based control strategy for demand responsive control of building thermal loads. The performance was assessed through comparisons to model predictive control and unconstrained reinforcement learning baselines with different comfort penalty factors. The test results show that the constrained

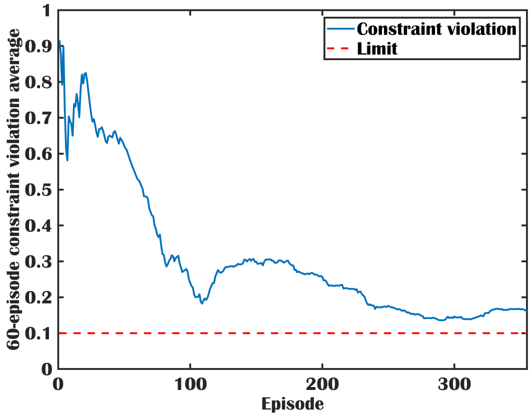

Fig. 6. Moving average constraint violation in CPO training.

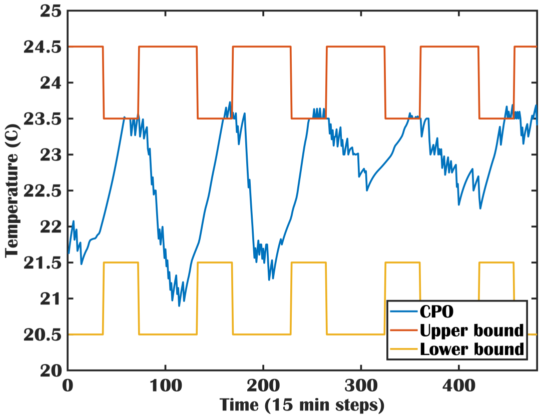

Fig. 7. Zone temperature under CPO strategy.

reinforcement learning-based strategy was able to reduce the electricity cost by 22%, relative to an energy-minimizing controller, with very minor temperature excursions out of the comfort zone, while the unconstrained reinforcement learning strategies led to either high energy costs or significant constraint violations, depending on the comfort penalty settings. The constrained reinforcement learning controller achieved performances very similar to the model predictive control benchmark, demonstrating its superior performance over unconstrained reinforcement learning techniques for building demand response. Further development is necessary to reduce the training time of these algorithms for field applications. Transfer learning and imitation learning are promising approaches to overcome this challenge and integrating them with constrained reinforcement learning algorithms will be addressed in future work.

TABLE II
Cost and Power Break-down

	Energy use (kWh)	Energy cost ($)	Cost savings (%)	Comfort violation (°C-step)
Emin	172.1	28.1	-	-
Umin	190.5	22.5	20	-
$\mathbf{TRPO_{0.01}}$	168	18.4	34.5	373
$\mathbf{TRPO_{0.1}}$	170	20.8	25.9	144
$\mathbf{TRPO_{1}}$	172	26.0	7.4	11
$\mathbf{PPO_{0.01}}$	159	17.9	36.3	451
$\mathbf{PPO_{0.1}}$	173	18.1	35.5	186
$\mathbf{PPO_{1}}$	166	25.7	8.1	16
CPO	210	21.7	22.7	12

References

[1] US Department of Energy. Benefits of demand response in electricity markets and recommendations for achieving them., 2006.

[2] Gregor P. Henze and Jobst Schoenmann. Evaluation of reinforcement learning control for thermal energy storage systems. *HVAC&R Research*, 9(3):259–275, 2003.

[3] Zhanhong Jiang, Michael J. Risbeck, Vish Ramamurti, Sugumar Murugesan, Jaume Amores, Chenlu Zhang, Young M. Lee, and Kirk H. Drees. Building hvac control with reinforcement learning for reduction of energy cost and demand charge. *Energy and Buildings*, 239:110833, 2021.

[4] Yan Du, Helia Zandi, Olivera Kotevska, Kuldeep Kurte, Jeffery Munk, Kadir Amasyali, Evan Mckee, and Fangxing Li. Intelligent multi-zone residential hvac control strategy based on deep reinforcement learning. *Applied Energy*, 281:116117, 2021.

[5] Liang Yu, Weiwei Xie, Di Xie, Yulong Zou, Dengyin Zhang, Zhixin Sun, Linghua Zhang, Yue Zhang, and Tao Jiang. Deep reinforcement learning for smart home energy management. *IEEE Internet of Things Journal*, 7(4):2751–2762, 2020.

[6] Kuldeep Kurte, Jeffrey Munk, Olivera Kotevska, Kadir Amasyali, Robert Smith, Evan McKee, Yan Du, Borui Cui, Teja Kuruganti, and Helia Zandi. Evaluating the adaptability of reinforcement learning based hvac control for residential houses. *Sustainability*, 12(18), 2020.

[7] Xiangyu Zhang, Dave Biagioni, Mengmeng Cai, Peter Graf, and Saifur Rahman. An edge-cloud integrated solution for buildings demand response using reinforcement learning. *IEEE Transactions on Smart Grid*, PP:1–1, 08 2020.

[8] Xiangyu Zhang, Rohit Chintala, Andrey Bernstein, Peter Graf, and Xin Jin. Grid-interactive multi-zone building control using reinforcement learning with global-local policy search, 2020.

[9] Donald Azuatalam, Wee-Lih Lee, Frits de Nijs, and Ariel Liebman. Reinforcement learning for whole-building hvac control and demand response. *Energy and AI*, 2:100020, 2020.

[10] Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[11] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015.

[12] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018.

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[14] *Constrained Markov Decision Processes*. 1999.

[15] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *CoRR*, abs/1705.10528, 2017.

[16] Jie Cai and James Braun. An inverse hygrothermal model for multi-zone buildings. *Journal of Building Performance Simulation*, 9(5):510–528, 2016.

[17] El Paso Electric. Small general service rate. https://www.epelectric.com/customers/rates-and-regulations/business-rates-and-information/texas-rate-tariffs-rules-and-regulations/texas-rate-tariffs, November 2021.

[18] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

[19] Gurobi solver. Online at https://www.gurobi.com/.

[20] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.