Global-Local Policy Search and Its Application in Grid-Interactive Building Control

Xiangyu Zhang, Yue Chen & Andrey Bernstein
National Renewable Energy Laboratory, Golden, CO 80401, USA
{xiangyu.zhang, yue.chen, andrey.bernstein}@nrel.gov

Abstract

As the buildings sector represents over 70% of the total U.S. electricity consumption, it offers a great amount of untapped demand-side resources to tackle many critical grid-side problems and improve the overall energy system's efficiency. To help make buildings grid-interactive, this paper proposes a *global-local policy search* method to train a reinforcement learning (RL) based controller which optimizes building operation during both normal hours and demand response (DR) events. Experiments on a simulated five-zone commercial building demonstrate that by adding a local fine-tuning stage to the evolution strategy policy training process, the control costs can be further reduced by 7.55% in unseen testing scenarios. Baseline comparison also indicates that the learned RL controller outperforms a pragmatic linear model predictive controller (MPC), while not requiring intensive online computation.

1 Introduction

With the acceleration and exacerbation of global climate change, human society is in dire need of technologies for decarbonization and sustainable development to avoid any irreversible consequences caused to the earth. Grid-interactive efficient building (GEB) control enables resources at grid edge to be harnessed, and with the provided flexibility, power systems can be operated in a cleaner manner with higher efficiency. Thus, GEB is gaining traction in recent years. Direct load control (DLC), see (San Diego Gas & Electric, 2023) for an example, allows utility companies to directly control customers' devices and is straightforward to implement. However, DLC does not explicitly consider specific building thermal condition and thus might jeopardize occupant comfort. Unlike DLC, MPC can combine building-centric and grid service objectives and achieves multi-objective optimal control, see (Drgoňa et al., 2020) for an extensive review. Despite the advantages of MPC, its massive deployment can be challenging. One of the reasons is that MPC requires on-demand computation to solve optimization problems during real time control (Zhang et al., 2021). Reinforcement learning (RL) policies, on the other hand, can be pre-trained offline and only require computationally cheap policy evaluation during real-time control. As a result, domain scientists are investigating using RL for optimal building control. For example, RL algorithms, including the deep Q-network (DQN) and asynchronous advantage actor-critic (A3C), are utilized for energy-saving while maintaining indoor comfort (Wei et al., 2017; Zhang et al., 2019). However, the discrete action spaces they employed usually require careful discretization to achieve good control performance and are more susceptible to the "curse of dimensionality" when applied to multi-zone control (Wei et al., 2017, Section 3.3). To use continuous action space, which greatly increases the policy search space and problem complexity, a Zap Q-learning method is leveraged in (Raman et al., 2020), though its application does not consider multi-zone building control. In addition, most prior building-RL studies are building-centric and do not consider enabling a building to be grid-interactive.

In this paper, we investigate developing an RL controller for the most complex single building control problem studied in RL-building literature. To achieve this, a global-local policy search method is proposed, which strategically combines merits of two different types of RL algorithm, to allow policy to converge to a better performing local optimum in the non-convex policy searching process. A full version of this paper is published in Zhang et al. (2022).

2 Global-local Policy Search

In deep RL, a policy network $\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$ parameterized by $\boldsymbol{\theta}$ is trained to maximize its control performance $J(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{a}_t \sim \pi_{\boldsymbol{\theta}}}(\sum_{t \in \mathcal{T}} \gamma^t r_t)$. The evolution strategies (ES) algorithm (Salimans et al., 2017) achieves this by maximizing a Gaussian smoothed version of the original objective:

$$V(\hat{\boldsymbol{\theta}}) := \mathbb{E}_{\theta \sim N(\hat{\boldsymbol{\theta}}, \sigma^2 I)} J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim N(0,I)} J(\hat{\boldsymbol{\theta}} + \sigma \boldsymbol{\epsilon}),$$

where $\boldsymbol{\theta}$ follows an isotropic multivariate Gaussian distribution with fixed covariance, i.e., $\boldsymbol{\theta} \sim N(\hat{\boldsymbol{\theta}}, \sigma^2 \mathbf{I})$, and $\hat{\boldsymbol{\theta}}$ is the mean parameter vector to be learned and σ is a standard deviation, which controls the smoothness. ES updates $\hat{\boldsymbol{\theta}}$ via $\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \alpha \nabla V(\hat{\boldsymbol{\theta}})$, and estimates $\nabla V(\hat{\boldsymbol{\theta}})$ using zero-order gradient estimation (ZOE) (Nesterov & Spokoiny, 2017):

$$\nabla V(\hat{\boldsymbol{\theta}}) \approx \frac{1}{\sigma} \mathbb{E}_{\boldsymbol{\epsilon} \sim N(0,I)} [\boldsymbol{\epsilon} \cdot J(\hat{\boldsymbol{\theta}} + \sigma \boldsymbol{\epsilon})]. \tag{1}$$

According to Salimans et al. (2017), without the need for backpropagation, ES is highly scalable and requires less computation per episode. In addition, by optimizing on a Gaussian smoothed surface, better properties than those of the original function are introduced, see (Nesterov & Spokoiny, 2017, Section 2) for more discussion. A direct benefit for this is the ability to converge to a better performing local optimum, if properly smoothed.

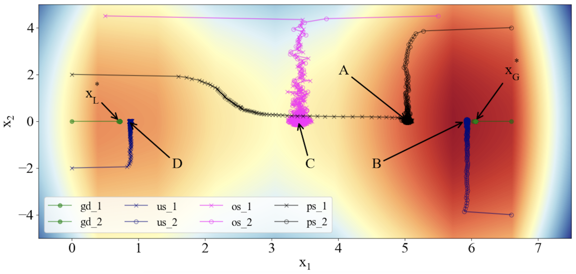


Figure 1: Eight learning trajectories on a non-convex function searching for a minimum.

Figure 1 shows an illustrative example of finding a minimum of a non-convex function $f(\mathbf{x})$, with two optima, i.e., $\mathbf{x}_L^*$ and $\mathbf{x}_G^*$, and $f(\mathbf{x}_G^*) < f(\mathbf{x}_L^*)$. Eight trajectories represented by four types of gradient descent (GD) convergence with *a)* accurate gradient $\nabla f(\mathbf{x})$ *b)* zero-order estimated gradient, i.e., (1), with small, medium and large values of σ, corresponding to "under-smooth (us)", "proper-smooth (ps)" and "over-smooth (os)" cases. Two trajectories, differentiated by the proximity of the initial point and an optimum, are generated for each cases. The comparison reveals that once properly smoothed (using large enough σ), the converged solution can escape the attraction of $\mathbf{x}_L^*$ and converging towards $\mathbf{x}_G^*$, even though the initial point is closer to $\mathbf{x}_L^*$, see "ps_1". Figure 2 provides 3D surfaces that explain this. However, also due to the function smoothing, the converged solution deviates from the true optimum and the deviation increases with the increment of σ, see 'B', 'A', 'C' vs. $\mathbf{x}_G^*$ in Figure 1. More details about this example are provided in Appendix A.

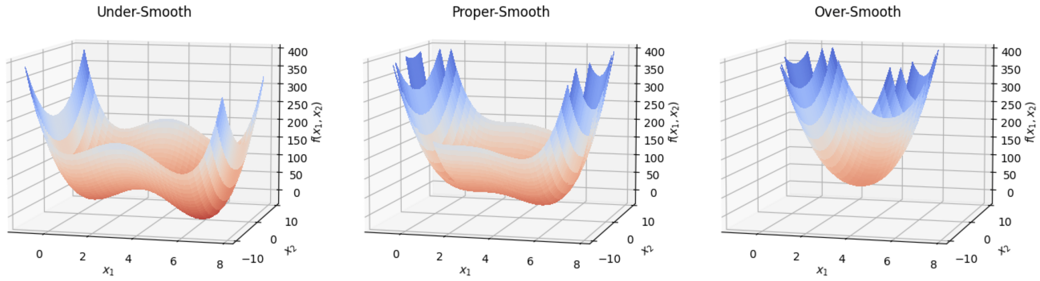


Figure 2: Three smoothed function surfaces with different σ.

To summarize, from this simple illustrative example, two key observations of the ZOE-based method's convergence feature are: *a)* An adequately large σ is needed to search with a more "global" vision and avoid poor-performing local optima; and *b)* Conversely, it requires σ to be small to converge "locally" to a better local optimum. Apparently, these two requirements are *conflicting*.

To reconcile this, a global-local policy search method is devised:

In the global search stage (1STG), the ES algorithm is used to search for a policy globally with a proper smoothing, and

In the local search stage (2STG) the ES pre-trained policy is refined locally using proximal policy optimization (PPO) (Schulman et al., 2017) to further push the policy closer to the "true optimum". PPO is suitable for fine-tuning because a) it directly optimizes on the original problem (instead of a smoothed objective) and b) the consideration of KL divergence in PPO policy update makes it suitable to improve an already good policy.

Finally, it is worth noting that the sub-optimal convergence is not discussed in the original ES paper. This is possibly because most RL benchmark problems are of the task-completion type, and with the ES learned sub-optimal policy, those tasks can still be completed anyway. However, when applying RL to real-world engineering problems, e.g., cost optimization, the policy improvement can have more practical meaning and thus provides additional incentive for conducting the 2STG fine-tuning.

3 Grid-Interactive Building Control

Consider a commercial building with $\mathcal{N} = \{1, ..., N\}$ thermal zones and a centralized air-conditioning (AC) unit in it. The AC can be controlled via two types of variables, i.e., chiller discharge air temperature $T^{da} \in [\underline{T}^{da}, \overline{T}^{da}]$ and zonal mass flow rate $\dot{m}^i \in [\underline{\dot{m}}^i_t, \overline{\dot{m}}^i_t], i \in \mathcal{N}$, to maintain indoor comfort. A grid-interactive building optimal control is formulated as follows by controlling $\mathbf{a}_t = [\dot{m}^1_t, \dot{m}^2_t, ..., \dot{m}^N_t, T^{da}_t]^\top \in \mathcal{A} \subset \mathbb{R}^{N+1}$:

$$\begin{aligned} \underset{\mathbf{a}_t \in \mathcal{A}, \forall t}{\text{minimize}} \quad & \sum_{t \in \mathcal{T}} \left[w_{(1,t)} \kappa_1 \sum_{i \in \mathcal{N}} \mathcal{D}(T^i_t) + w_{(2,t)} \kappa_2 p_t(\mathbf{a}_t) \Delta t + w_{(3,t)} \kappa_3 ((p_t(\mathbf{a}_t) - \overline{p}_t)^+)^2 \right] \\ \text{subject to} \quad & \mathbf{T}_t = \mathcal{F}\left(\mathbf{T}_{t-1}, \mathbf{a}_t, \boldsymbol{\epsilon}_t\right) \quad (\forall t \in \mathcal{T}), \end{aligned} \tag{2}$$

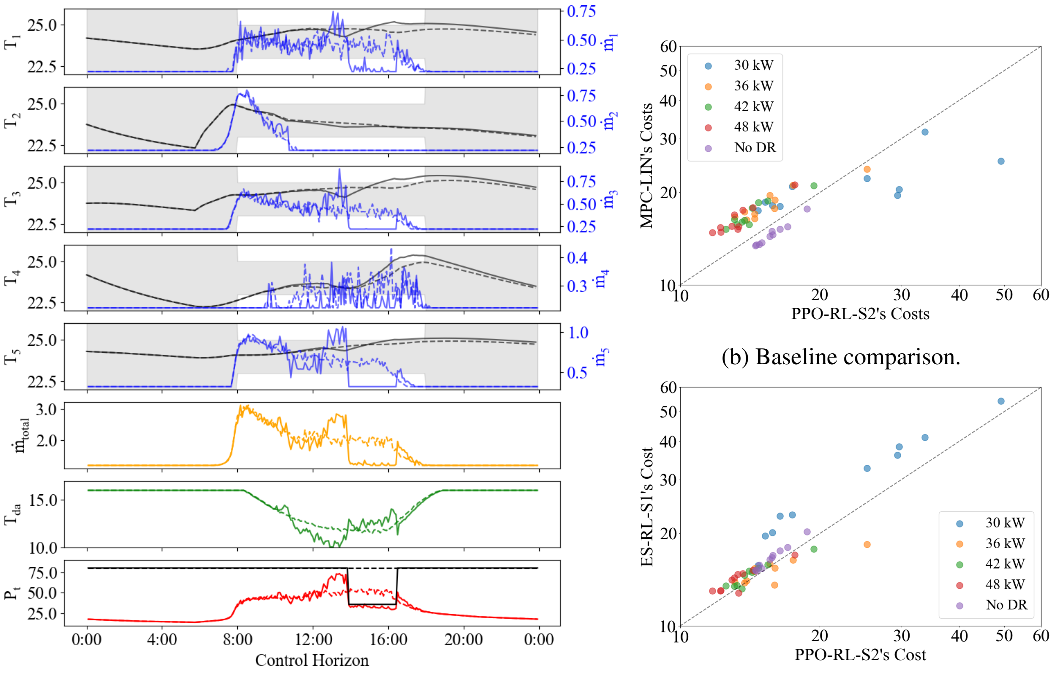
where $w_{(i,t)}$ and κ_i are weighting factors and monetizing factors for the three objectives to be minimized, i.e., costs associated with thermal discomfort $\mathcal{D}$, energy consumption $p_t(\mathbf{a}_t)\Delta t$ and power limit violation $(p_t(\mathbf{a}_t) - \overline{p}_t)^+$. In (2), $(\cdot)^+ = \max(0, \cdot)$, $p_t(\mathbf{a}_t)$ calculates AC power consumption at step t, and $\overline{p}_t$ is the demand response (DR) power limit given by the utility company. Zonal temperature $\mathbf{T}_t$ is determined by the building thermal dynamics $\mathcal{F}$, and $\boldsymbol{\epsilon}_t$ denotes environmental disturbances. Problem (2) is formulated as a Markov Decision Process, with more details in Appendix B, and a policy $\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$ is trained to implement optimal control.

4 Results

The global-local policy search method is used to train $\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)$ for (2), using one month of environmental disturbance data for training and use the next ten days (unseen scenarios) for testing. DR events are generated randomly, i.e., if or not there will be a DR event, duration of the DR event and what are the power limit values $\overline{p}_t$ for $t \in \mathcal{T}$. The trained control policy needs to able to handle all these scenarios.

4.1 Control Efficacy

To examine the control behavior of the trained controller, Figure (3a) shows the control trajectories of the trained controller in one testing scenario, with two cases: with a DR event and without an event. It can be seem that the demonstrated control behavior is desirable: i) zonal temperature are mostly kept within the comfort band over $\mathcal{T}$; ii) DR power limits are satisfied; iii) in the DR case, proactive prior-event control is observed to prepare the building for the in-coming DR event; iv) though not instructed, the policy learned that Zone 2 (an east-facing zone) does not require pre-cooling prior to the DR event; and v) all cooling air goes to Zone 4 (the west-facing zone) during the DR event to counter the thermal discomfort. In addition to inspect the behavior of the learned control policy in one specific scenario, we also compared the learned control policy with an optimization based controller under multiple DR and weather scenarios as well, see Figure (3b). Over these testing scenarios and compared with the linear MPC, the two-stage trained RL controller reduces average control costs by 4.16%.



(a) Single testing scenario rollout.

(b) Baseline comparison.

(c) Fine-tuning comparison.

Figure 3: Testing the global-locally searched RL policy in testing scenarios. (a) In all sub-figures, *dashed* lines are for the case without a DR event and the *solid* ones are for the DR scenario. Shaded areas in the first five sub-figures reveal the temperature comfort band and black lines (both dashed and solid) in the bottom figure represent the DR power limit $\overline{P}_t$. (b) Cost comparison with a linear MPC baseline, each dot represents one of the fifty testing scenarios. Most of these dots are above to the dashed line, and on average RL controller can reduce costs by 4.16%, when compared with the baseline. (c) Cost comparison with the ES pre-trained policy.

4.2 Benefit for local fine-tuning

Table 1: 2STG Cost Reduction.

σ	Converged Episodic Cost		
	1STG	2STG	δ (%)
0.01	18.74	14.48	22.73%
0.02	15.67	14.55	7.15%
0.05	15.09	14.17	6.49%

Warmstarted with the 1STG ES pre-trained policy, PPO is used for policy fine-tuning in 2STG. Table 1 shows how much improvement, denoted as δ in percentage, can be achieved in scenarios where three different smoothing factors are used in 1STG. In addition to training, the performance comparison of 1STG ES pre-trained and 2STG PPO fine-tuned controllers are shown in Figure (3c). Over these 50 testing scenarios, the 2STG local fine-tuning can help achieve 7.55% cost reduction.

5 Conclusion

In this paper, we proposed a global-local policy search method, which first use a ZOE-based method to search globally and escape from the attraction of poor performing local optima, and then fine-tunes the policy using policy gradient method. The effectiveness and advantages of the proposed method were demonstrated in a multi-zone grid-interactive building control problem. We hope our findings can provide some insights on using RL for grid-interactive building control, enabling more buildings to provide grid services through DR programs and collectively contribute to a cleaner and more efficient energy system.

Acknowledgements

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

Sourav Dey, Thibault Marzullo, Xiangyu Zhang, and Gregor Henze. Reinforcement learning building control approach harnessing imitation learning. *Energy and AI*, 14:100255, 2023.

Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, Juraj Oravec, Michael Wetter, Draguna L Vrabie, et al. All you need to know about model predictive control for buildings. *Annual Reviews in Control*, 50:190–232, Sep. 2020.

Yang Liu, Nanpeng Yu, Wei Wang, Xiaohong Guan, Zhanbo Xu, Bing Dong, and Ting Liu. Coordinating the operations of smart buildings in smart grids. *Applied Energy*, 228:2510–2525, Oct. 2018.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, Nov. 2017.

June Young Park, Mohamed M Ouf, Burak Gunay, Yuzhen Peng, William O'Brien, Mikkel Baun Kjærgaard, and Zoltan Nagy. A critical review of field implementations of occupant-centric building controls. *Building and Environment*, 165:106351, 2019.

Matias Quintana, Zoltan Nagy, Federico Tartarini, Stefano Schiavon, and Clayton Miller. Comfortlearn: enabling agent-based occupant-centric building controls. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pp. 475–478, 2022.

Naren Srivaths Raman, Adithya M Devraj, Prabir Barooah, and Sean P Meyn. Reinforcement Learning for Control of Building HVAC Systems. In *2020 American Control Conference (ACC)*, pp. 2326–2332. IEEE, Jul.1-3, 2020.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

San Diego Gas & Electric. AC Saver (Summer Saver), 2023. Accessed: Feb. 3rd, 2023. [Online]. Available: https://www.sdge.com/residential/savings-center/rebates/your-heating-cooling-systems/summer-saver-program.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Olli Seppanen, William J Fisk, and David Faulkner. Cost benefit analysis of the night-time ventilative cooling in office building. 2003.

Tianshu Wei, Yanzhi Wang, and Qi Zhu. Deep reinforcement learning for building HVAC control. In *Proceedings of the 54th Annual Design Automation Conference*, pp. 1–6, Jun.18, 2017.

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*, 12(1):420–431, Jan. 2021. doi: 10.1109/TSG.2020.3014055.

Xiangyu Zhang, Yue Chen, Andrey Bernstein, Rohit Chintala, Peter Graf, Xin Jin, and David Biagioni. Two-stage reinforcement learning policy search for grid-interactive building control. *IEEE Transactions on Smart Grid*, 13(3):1976–1987, 2022.

Zhiang Zhang, Adrian Chong, Yuqi Pan, Chenlu Zhang, and Khee Poh Lam. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. *Energy and Buildings*, 199:472–490, Sep. 2019.

A Numerical Experiment Details

The non-convex function used in this numerical example is $f(\mathbf{x}) = 20 - 70x_1 + 65.7x_1^2 - 17.1x_1^3 + 1.3x_1^4 + 1.6x_2^2$. Table 2 shows the smoothing factors and initial points used in the experiment in Section 2.

Table 2: ZOE Trial Parameters

Type	σ	**Trial (Initial Points)**
Under Smoothed	0.5	us_1 (0.0, -2.0), us_2 (6.6, -4.0)
Properly Smoothed	1.285	ps_1 (0.0, 2.0), ps_2 (6.6, 4.0)
Over Smoothed	2.5	os_1 (0.5, 4.5), os_2 (5.5, 4.5)

B Additional Details on Grid-Interactive Building Control

B.1 Term definitions

According to Seppanen et al. (2003), higher temperature can cause reduction in occupants' productivity; and Dey et al. (2023) further monetizes such comfort and includes it in the objective function. Without loss of generality, in this study, the cost associated with zonal thermal discomfort $\mathcal{D}(T_t^i)$ in (2) is defined as the temperature deviation from a pre-defined comfort band $[\underline{T}^i, \overline{T}^i]$ with a piece-wise function:

$$\mathcal{D}(T_t^i) := \begin{cases} \max(T_t^i - \overline{T}^i, (T_t^i - \overline{T}^i)^2) & (T_t^i > \overline{T}^i) \\ \max(\underline{T}^i - T_t^i, (\underline{T}^i - T_t^i)^2) & (T_t^i < \underline{T}^i) \\ 0.0 & (\text{else}) \end{cases} \tag{B.1}$$

where T_t^i is the indoor temperature of zone i at step t. See Park et al. (2019); Quintana et al. (2022) for a more occupant-centric building control performance index.

The AC power consumption $P(\mathbf{a}_t, T_t^{out})$ is given by:

$$P(\mathbf{a}_t, T_t^{out}) := a(T_t^{out} - T_t^{da}) \sum_{i=1}^{N} \dot{m}_t^i + b(\sum_{i=1}^{N} \dot{m}_t^i)^3 + c. \tag{B.2}$$

The first term in (B.2) describes the chiller power (Liu et al., 2018) and the rest depicts the fan power; T_t^{out} is the outdoor temperature and a, b and c are known constants. Note, the chiller power term has products of $T_t^{da}\dot{m}_t^i$, both are decision variables, this bilinear term makes the problem nonlinear.

B.2 MDP Formulation

The optimal control problem depicted by (2) is formulated into an MDP with the following elements:

State: To properly guide the RL controller in decision-making, the state representation typically contains information regarding the current system status and other information related to its future evolving trajectory. As a result, we define the state in this study as

$$\mathbf{s}_t := [\mathbf{T}_t, \mathbf{T}_{t,-K}^{out}, \omega, \sin_t, \cos_t, t, \overline{\mathbf{p}}_{t,K}, \mathbf{w}_t]^\top,$$

including

1. zonal temperature $\mathbf{T}_t$ reflecting current status,
2. outdoor temperature for the last K steps $\mathbf{T}_{t,-K}^{out}$ implying weather condition,
3. weekday indicator ω, trigonometric encoding of time $\sin_t, \cos_t$ reflecting the occupancy schedule, control step number t indicating the control progress,
4. DR signal received from the utility company, i.e., power limit for the next K steps $\overline{\mathbf{p}}_{t,K}$,
5. objectives' weighting factors $\mathbf{w}_t$, provided by the building operator on how to balance the objectives of building thermal condition, energy consumption and grid-service.

Action: RL control action is the same as the decision variables in (2) as $\mathbf{a}_t = [\dot{m}_t^1, \dot{m}_t^2, ..., \dot{m}_t^N, T_t^{da}]^\top \in \mathcal{A} \subset \mathbb{R}^{N+1}$.

Reward: The reward is naturally defined as the negative single step cost in (2), i.e., $r_t = -[w_{(1,t)}\kappa_1 \sum_{i\in\mathcal{N}} \mathcal{D}(T_t^i) + w_{(2,t)}\kappa_2 p_t(\mathbf{a}_t)\Delta t + w_{(3,t)}\kappa_3((p_t(\mathbf{a}_t) - \overline{p}_t)^+)^2]$.

State transition: The state transition is determined by the building thermal dynamics $\mathbf{T}_t = \mathcal{F}\left(\mathbf{T}_{t-1}, \mathbf{a}_t, \boldsymbol{\epsilon}_t\right)$ and environmental disturbances.