Online Comfort-Constrained HVAC Control via Feature Transfer

Jing Yu
California Institute of Technology
University of Washington
USA
jing5@uw.edu

Tianyu Zhang
Autodesk Research
Canada
tianyu.zhang@autodesk.com

Omid Ardakanian
University of Alberta
Canada
oardakan@ualberta.ca

Adam Wierman
California Institute of Technology
USA
adamw@caltech.edu

ABSTRACT

Policy transfer can reduce energy consumption of the Heating, Ventilation, and Air Conditioning (HVAC) system in a target building without requiring extensive offline data or costly online interaction. However, the energy savings often come at the expense of violating constraints, such as thermal comfort constraints, in the target building. In this paper, we propose a novel approach to augment black-box policy transfer methods for simultaneous minimization of energy consumption and thermal comfort constraint violation. To ensure compliance with these constraints, the algorithm leverages nonlinear feature functions learned from historical data of source buildings and adapt these features to approximate the target building dynamics via online learning in a sample efficient manner. We present an upper bound on the number of time steps during which thermal comfort violations may occur under worst-case disturbances, and also establish a finite stopping time beyond which all constraints are consistently met, assuming certain stochastic properties of disturbances. Additionally, we evaluate the algorithm using the EnergyPlus simulator across 19 climate locations in January and July. OCFT reduces the total number of constraint violation of the black-box policy transfer algorithm by an average of 81.28% while resulting in 11.23% more energy consumption on average.

CCS CONCEPTS

• **Theory of computation** → **Online learning algorithms**; • **Computing methodologies** → *Reinforcement learning*.

KEYWORDS

HVAC Energy Optimization

ACM Reference Format:
Jing Yu, Tianyu Zhang, Omid Ardakanian, and Adam Wierman. 2025. Online Comfort-Constrained HVAC Control via Feature Transfer. In *The 16th ACM International Conference on Future and Sustainable Energy Systems (e-Energy '25), June 17–20, 2025, Rotterdam, Netherlands.* ACM, New York, NY, USA, 14 pages. https://doi.org/xxxx

1 INTRODUCTION

Heating, Ventilation, and Air Conditioning (HVAC) systems are among the largest energy consumers in commercial buildings, accounting for roughly 40% to 60% of the total energy use of the building. As the push to decarbonize the building stock gains momentum, optimizing the performance of HVAC systems becomes a critical priority. Currently, most HVAC systems are controlled using rule-based controllers. These reactive controllers satisfy thermal comfort requirements by keeping room temperature within bounds during the specified occupancy events, but do not optimize energy consumption.

In recent years, machine learning has shown great potential to improve HVAC energy efficiency and enable dynamic adaptation to changing building conditions and occupancy patterns [49]. Despite significant progress in machine learning-based control strategies, it is well known that training control policies is data-intensive and computationally resource-intensive. Moreover, due to the heterogeneity and large scale of commercial buildings in diverse climate conditions, new models and policies must be learned for each novel building, rendering machine learning-based approaches prohibitively expensive and hard to scale for real-world application.

A promising approach to address the challenge of scalability is transfer learning, especially in the domain of reinforcement learning [46]. The core idea of transfer learning is that policies learned in one task can fast-track the learning for a related yet distinct task. Recently, work has begun to apply various transfer learning techniques in HVAC control, such as combining transfer learning and imitation learning [16], and promoting diversity in a population of agents for transfer learning [28]. However, a major drawback of these transfer learning methods for HVAC control is that most methods disregard individual thermal comfort requirements and optimize energy efficiency at the expense of large violations of comfort constraints. This is because comfort constraints are inherently specific to each building and cannot be considered during policy learning on the source buildings. Given the importance of thermal comfort and its impact on the well-being and productivity of building occupants [29], it is imperative to design HVAC control strategies that can quickly learn to minimize energy consumption while simultaneously adhering to thermal comfort and other operational constraints.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

e-Energy '25, June 17–20, 2025, Rotterdam, Netherlands

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN xxxx…$15.00
https://doi.org/xxxx

Contributions. We propose Online comfort-constrained Control via Feature Transfer (OCFT), an algorithm that augments black-box transfer learning methods in order to simultaneously minimize energy consumption and thermal comfort constraint violations while respecting hard operational constraints (Algorithm 1). Given a black-box transfer learning policy, OCFT follows actions suggested by the policy to minimize energy consumption when constraints are not active and generates actions that adhere to thermal comfort bounds when the transferred policy would have violated them.

The high-level idea of OCFT is shown in Figure 1, where the algorithm leverages nonlinear feature functions that capture zone-level dynamics, which are learned using historical data from source buildings. OCFT adapts these features to the target building online, with the goal of approximating the target building zone-level dynamics in a sample-efficient manner. Motivated by recent advances in online learning and control of linear systems, we perform nonlinear feature adaptation using online nested convex body chasing algorithms. Given the learned approximate dynamics, the algorithm then augments an arbitrary black-box policy by solving a robust optimization problem to select actions that closely track actions suggested by the black-box policy while conforming to operational and thermal comfort constraints.

We present theoretical insights for the design of OCFT in both stochastic and worst-case settings. When the disturbances are only assumed to be bounded, but can be arbitrary and potentially adversarial, Theorem 1 guarantees finite thermal constraint violations of OCFT during the online adaptation process. If the disturbances have additional stochastic properties, we further have Corollary 1 guaranteeing that OCFT will not violate any thermal comfort constraints after a finite time. The main contribution of this paper is the extensive evaluation on the performance of the proposed algorithm. Compared to the default air system controller in EnergyPlus, three state-of-the-art RL policy transfer methods, and a baseline controller that uses least squares estimation for parameter adaptation, OCFT achieves the lowest thermal comfort violation rate in all 19 climate locations as designated by the U.S. Department of Energy simulated in peak heating and cooling seasons.

Related Work. Designing learning-based control strategies for the HVAC system is a well-studied topic. Previous work in this area can be broadly classified into three categories (a) learning-based Model Predictive Control (MPC) [20], where an approximate system dynamics model [7] or the cost function and constraints [15] are learned from data, (b) reinforcement learning (RL), where a policy is learned using previously collected data [33, 34, 56] and/or through online interaction with the building environment [14, 38, 58], and (c) transfer learning, where the policy learned on a source building is transferred to the target building and adapted to the new environment in an online fashion [16, 21, 28, 51].

The performance of learning-based MPC and model-based RL [3, 13, 19] strategies highly depends on the accuracy of the dynamics model. Considering the diversity and the scale of large commercial buildings, as well as the difficulty of learning a sufficiently accurate model from passively collected data, these strategies often struggle to achieve acceptable performance in practice.

Model-free RL strategies can achieve high performance without relying on a model, but they need extensive offline data or an extremely large amount of interaction to learn a near-optimal policy [58], which is costly and potentially dangerous, hence not affordable in real buildings. Learning an RL policy through interactions with a building in simulation also requires access to a high-fidelity building simulator (e.g. EnergyPlus) and an accurate model of the target building; otherwise, the learned policy may perform poorly in the real building due to the sim-to-real gap [6]. Learning candidate RL policies on a variety of buildings, followed by policy selection based on a small amount of historical data from the target building and transfer of the best policy to the target building addresses the drawbacks of the other approaches. It is shown to be capable of achieving higher performance than the existing rule-based controller and an RL policy that is learned from scratch through interaction with the target building only [28]. However, deploying the transferred policy to the target building may lead to a high constraint violation rate. This is because the target building's constraints are not imposed during policy learning on source buildings, as constraints are usually specific to each building. Overall, previous work on control policy transfer does not provide theoretical guarantees for constraint satisfaction in the target building. Constraint satisfaction is merely *encouraged* via reward shaping [16, 21, 51]. Another approach is to use a constraint-conforming controller in tandem with the RL controller [40], but this backup controller must be designed specifically for the target environment.

In recent years, augmenting a black-box control policy with safety guarantees in the form of constraint satisfaction has received significant attention, although it has not been applied in the HVAC control domain. Notably, a Model Predictive Safety Filter (MPSF) has been proposed in [30, 48] to compute a safe backup trajectory by minimizing modifications to the potentially unsafe control input, given an approximate system dynamics model. Although this approach can prevent constraint violation, interventions of the safety filter would result in performance degradation of the control policy. Our approach differs from MPSF in that we work with nonlinear models and ensure that the transferred policy satisfies all constraints in the target building after a finite number of constraint violations while minimizing the performance degradation resulted from constraint conformation.

Another related line of research is safe transfer RL, e.g., [22, 52, 53]. Algorithmically, to ensure constraint satisfaction, these approaches commonly formulate constraints as cost functions under the constrained Markov decision process (CMDP) framework. In particular, constraints are cast as upper bounding the *expectation* of constraint functions. In contrast, our results provide an explicit *worst-case number of steps in which constraints are violated.* Moreover, our guarantee is online, whereas safety-constrained policy transfer methods only guarantee constraint satisfaction on average, once the algorithms converge. These methods are complementary to OCFT, since OCFT is an augmentation approach that safeguards any black-box policies, including safe transfer RL policies, with worst-case guarantees.

2 MODEL

We consider an unknown target HVAC system with a parameterized zone-level model of the following form,

$$s_{t+1} = \theta^{\star\top}\Phi(s_t, z_t, a_t) + d_t \tag{1}$$

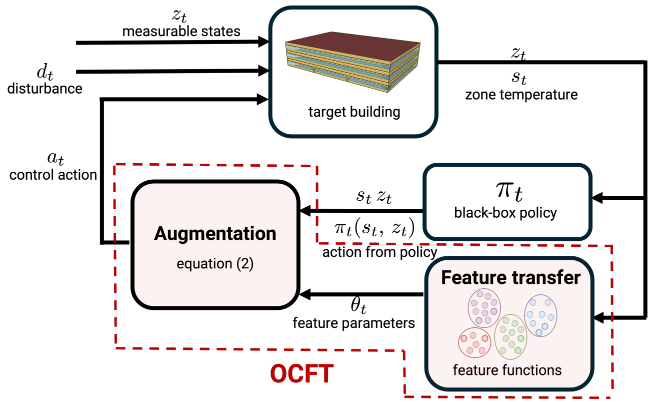

Figure 1: Overview of the Online comfort-constrained Control via Feature Transfer (OCFT) framework

where $s_t \in \mathbb{R}$ is the zone temperature, $z_t \in \mathbb{R}^n$ includes the measurable states such as occupancy level and outdoor temperature, $a_t \in \mathbb{R}^m$ is the zone-level action such as damper position, and $d_t \in \mathbb{R}$ denotes exogenous inputs that capture unmodeled dynamics and disturbances such as thermal interactions between zones. The function $\Phi : \mathbb{R}^{n+m+1} \to \mathbb{R}^q$ is a vector of q nonlinear feature functions that model the dynamics of the zone-level system. The zone-level dynamics (1) is parameterized by the unknown coefficient $\theta^\star \in \mathbb{R}^q$, which we will refer to as the true model parameter. The parameterization is popular in nonlinear system modeling, e.g., [12, 13, 37]. It generalizes the linear time-invariant RC dynamics commonly used in model-based control for HVAC systems [17, 36, 44], where Φ is specialized to linear functions.

Although commercial buildings can have diverse building-level configurations with different cooling and heating modules, zone-level thermal dynamics for commercial buildings are likely to share features. Such features can be learned using historical data from a variety of source buildings. Our algorithm leverages these similarities for fast adaptation in novel buildings by updating the model parameter θ online as the controller collects sequential data generated by the HVAC system in the target building. For ease of notation, we assume all zones have the same dimension for states and actions, as well as the same number of nonlinear feature functions. The generalization to heterogeneous dimensions is straightforward.

Remark 1 (Quality of features). *The dynamical model (1) requires the offline learned feature functions Φ to be diverse enough to subsume the dynamical behavior of the target building zones. In the experiments (Figure 7), we observe that OCFT has inherent robustness against a small number of feature functions. However, if the features are very low quality, i.e., if features are all learned from a single zone in one season and the proposed algorithm is deployed on various zones in a different season, the performance might deteriorate.*

The fundamental goal of HVAC control is to compute a_t in order to meet the thermal comfort constraints while minimizing energy consumption. In this paper, we model the thermal comfort constraint with $\underline{T}$ and $\overline{T}$ denoting the lower and upper zone temperature limit, respectively. For the ease of exposition in the analysis, we let the thermal comfort constraints be time-independent. We use season- and occupancy-dependent thermal constraints in all the numerical experiments in Section 5.2. For the parameterized zone-level model, we make the following assumptions.

Assumption 1 (Bounded states, actions, features, and parameters). *The zone temperature s, measurable state z and action a belong to compact and convex sets $\mathcal{S}$, $\mathcal{Z}$, and $\mathcal{A}$ respectively. Further, the nonlinear feature functions Φ are convex and bounded such that for all $s \in \mathcal{S}$, $z \in \mathcal{Z}$, and $a \in \mathcal{A}$,*

$$\|\Phi(s, z, a)\|_2 \le T_{\max}$$

for $T_{\max} \in \mathbb{R}$. Moreover, the true parameter $\theta^\star$ belongs to a known compact and convex uncertainty set Θ.

Assumption 2 (Bounded disturbances and parameters). *For all zones, the disturbances d_t belong to a hypercube such that $\|d_t\|_\infty \le W$ for all $t \ge 0$ with a known constant $W > 0$.*

In the real world, the zone temperatures, measurable states, e.g., outdoor temperature, occupancy, etc., and exogenous disturbances are always bounded. Zone-level actions, like damper positions, are inherently bounded and may also be subject to additional operational constraints, such as ramping limits. Assumption 1 also states that the outputs of the feature functions are bounded in magnitude by $T_{\max}$. This is a mild assumption that ensures the feature functions make reasonable zone temperature predictions. The convexity requirement in Assumption 1 is satisfied for common choices of feature functions such as linear dynamics and deep neural networks in building applications, e.g., input convex neural networks [15]. The initial uncertainty set Θ in Assumption 1 enables incorporating any prior information about the nonlinear features, such as a bounded region around some initial estimation. If there is no prior information, one can use $\Theta = \{\theta : \|\theta\|_2 \le \kappa\}$, which is convex and compact, with the knowledge that the true parameter is bounded by some constant κ. Our final assumption is on the robust feasibility of the HVAC control problem under comfort constraints:

Assumption 3 (Robust feasibility). *There exists a known constant $\epsilon > 0$ and for all parameters $\theta \in \Theta$, states $s \in \mathcal{S}$, $z \in \mathcal{Z}$, there exists an action $a' \in \mathcal{A}$ such that*

$$\overline{T} + (W + \epsilon) \le \theta^\top \Phi(s, z, a') \le \underline{T} - (W + \epsilon).$$

where $\underline{T}$ and $\overline{T}$ are respectively the lower and upper thermal comfort level.

Intuitively, Assumption 3 ensures the existence of a feasible action that will robustly satisfy the thermal comfort constraints with a robustness margin of $\epsilon > 0$ despite disturbances, which satisfy Assumption 2. .

3 ALGORITHM

We propose a novel algorithm for online comfort-constrained HVAC control via feature transfer (OCFT), which is summarized in Algorithm 1. Before online deployment, the algorithm uses a small amount of log data collected with a default rule-based controller in the target building to initialize the black-box policy transfer learning algorithm and construct a model uncertainty set based on log data (line 1). During deployment, the algorithm updates the model uncertainty set based on the sequentially revealed online data (line 4) and leverages nested convex body chasing (CBC) algorithms to select hypothesis models from the latest model uncertainty set. The selected hypothesis model in each step is used as part of a robust optimization problem that augments the black-box transfer learning policy so that thermal constraints are satisfied. If the transfer learning method requires online updates (i.e., adaptation), then the algorithm will use the updated policy in the next iteration.

Algorithm 1 Online Comfort-constrained HVAC Control (OCFT)

```
Input: disturbance bound $W$, feature functions $\Phi$, log data $\mathcal{D}$
Algorithm parameter: trust parameter $\lambda$, weights $\eta_1, \eta_2$
1: Generate $\mathcal{P}_0$ and initialize $\pi_0$ with log data        ▷ Warm start
2: for $t = 1, 2, \ldots$ do
3:     Observe $s(t)$
4:     Construct consistent model set $\mathcal{P}_t$ with $\mathcal{P}_t :=$
           $\{\theta : \|s_t - \theta^\top \Phi(s_{t-1}, z_{t-1}, a_{t-1})\|_\infty \le W\} \cap \mathcal{P}_{t-1}$
5:     $\theta_t \leftarrow \mathrm{NCBC}(\mathcal{P}_t)$        ▷ Online model adaptation
6:     $a_t \leftarrow \mathrm{AUG}\,(s_t\,;\, \theta_t,\, \pi_t)$ with (2)        ▷ Policy augmentation
7:     $\pi_{t+1} \leftarrow \mathrm{RL\text{-}TRANSFER}(s_t, \pi_t)$        ▷ Policy update (optional)
```

3.1 Warm start

Given a transfer learning algorithm RL-TRANSFER and log data $\mathcal{D} = \{s_k,\, z_k,\, a_k\}_{k=0}^{T_0}$ with horizon T_0, the algorithm initializes the black-box transfer learning policy as π_0 after training on log data. Generally, the amount of log data is considerably less than that needed to retrain a new policy. In our experiments, we use 2 weeks of log data collected using a default rule-based controller sampled at 15-minute interval with $T_0 = 1344$. In contrast, learning a high-quality policy from scratch would require many months of data from the target building.

We will also construct a *consistent model set* $\mathcal{P}_0$ based on the log data by refining the known uncertainty set Θ in Assumption 2 as follows:

$$\mathcal{P}_0 = \{\theta \in \Theta : \|s_{k+1} - \theta^\top \Phi(s_k, z_k, a_k)\|_\infty \le W\,, \\ \forall\, (s_k, z_k, a_k) \in \mathcal{D}\}.$$

The set $\mathcal{P}_0$ contains all model parameters that are consistent with the temperature dynamics observed in the log data. In other words, for all $\theta \in \mathcal{P}_0$, there exists a sequence of admissible disturbances $(\hat{d}_0, \hat{d}_1, \ldots, \hat{d}_{T_0-1})$ that satisfy Assumption 2 such that the log data can be generated according to (1). The log data helps eliminate implausible model parameters in Θ and reduce the uncertainty of the model on the target building.

3.2 Online model adaption

During online operation, the algorithm continuously updates the consistent model set with the new observations (line 4) using the same procedure as in the warm start period while the temperature transitions are sequentially revealed (line 3). The set $\mathcal{P}_t$ is crucial for guaranteeing the satisfaction of the comfort constraints, as it keeps track of the uncertainty about the HVAC dynamics model. We note that $\mathcal{P}_t$ is nested, i.e., $\mathcal{P}_t \subseteq \mathcal{P}_{t-1}$ by definition, which reflects the reduction in uncertainty as more online data is observed. Intuitively, the smaller $\mathcal{P}_t$, the more knowledge OCFT has about the target building, resulting in less comfort constraint violation.

Why not use other system identification techniques? The set $\mathcal{P}_t$ is sometimes referred to as the membership set, e.g., [1, 8, 35] in the system identification literature. Even in the idealized setting where disturbances are i.i.d. stochastic and Φ is linear, [31, 50] showed that $\mathcal{P}_t$ converges to the true system parameter significantly faster than the best estimator for linear systems. In particular, $\mathcal{P}_t$ converges linearly to the true parameter with respect to number of samples (estimation error scales as $O(1/t)$). In contrast, the least squares estimator, which has been shown to achieve the information theoretical lower bound of the sample complexity for model learning in linear systems, converges to the true parameter much slower (estimation error scales as $O(1/\sqrt{t})$). The set membership estimation method breaks through the lower bound by taking advantage of the knowledge that the disturbances are *bounded* as in Assumption 2. In HVAC systems, the states, actions, and disturbances are naturally bounded, rendering the set membership estimation an ideal sample-efficient model uncertainty estimation method. Our usage of $\mathcal{P}_t$ is therefore motivated by these theoretical findings in linear systems. In Section 4, we show that indeed the same sample complexity holds for the nonlinear feature model (1).

Following the construction of $\mathcal{P}_t$, we select a *hypothesis model*, which will be used for the downstream augmentation task (line 5). The hypothesis model is important for balancing generating "safe" actions that satisfy constraints and generating "exploratory" actions that may violate constraints but will reduce the size of $\mathcal{P}_t$. In particular, we use the nested convex body chasing algorithms for such a model selection (line 5).

Nested CBC is an online learning problem, where in every round $t \ge 0$, a player is presented a convex set $\mathcal{K}_t \subset \mathbb{R}^n$. The player then selects a point $q_t \in \mathcal{K}_t$ with the objective of minimizing the cost defined as the total movement of the selection for T rounds, *e.g.*, $\sum_{t=1}^{T} \|q_t - q_{t-1}\|$ for a given initial condition $q_0 \notin \mathcal{K}_1$. In our problem setting, the consistent model sets $\mathcal{P}_t$ is the sequentially revealed convex sets while the selected points are the hypothesis models θ_t. Nested CBC is a well studied problem, with many algorithms [5, 10, 11, 43] that trade off computational complexity and competitive ratio guarantees, where the cost incurred by the algorithm is at most a constant factor away from the cost incurred by the offline optimal algorithm that has the knowledge of the entire sequence of the sets in the worst case. We leave the choice of the nested CBC algorithm to the user and denote it as $\mathrm{NCBC}(\cdot)$.

Why convex body chasing? While there are many potential alternatives to select the hypothesis point in $\mathcal{P}_t$, it has been shown that using nested CBC to select models for downstream online control tasks, e.g., [25, 55], induces an endogenous trade-off between exploration and exploitation of the model uncertainty set. In particular, nested CBC helps select models that produce control inputs that *either* reduce the size of the uncertainty set (exploration) but may violate thermal comfort constraints, *or* produce constraint-conforming actions despite disturbances and the model uncertainty (exploitation) but do not reduce model uncertainty. Intuitively, such a trade-off limits the *total amount of violation* of the thermal comfort constraints during the online learning process before the model uncertainty set becomes small enough for robust constraint satisfaction. We formalize this intuition in Section 4, where we provide a worst-case constraint violation bound for Algorithm 1.

Computational complexity of $\mathcal{P}_t$ and nested CBC. The consistent model set $\mathcal{P}_t$ can be represented by $2(T_0+t)$ linear constraints since each transition data $(s_k, s_{k-1}, z_{k-1}, a_{k-1})$ observed up until t provides a linear constraint on θ_t under the upper and lower bound of admissible disturbances. However, the number of such constraints increases linearly with t. To address this, many computationally efficient approaches have been proposed based on approximations of the set $\mathcal{P}_t$, e.g., [8, 35] and constraint sampling, e.g., [54]. Among all nested CBC algorithms, the Steiner point selection [43] achieves the optimal competitive ratio and can be solved using randomized linear programs [5]. However, the number of randomized linear programs increases quadratically with the dimension of the selected points, which can pose challenges if the number of feature functions in (1) is large. A common alternative is to use projection, where the algorithm finds the closest point with respect to Euclidean distance in the new convex set from the previously selected point, i.e., $\theta_t = \operatorname{argmin}_{\theta \in \mathcal{P}_t} \|\theta - \theta_{t-1}\|_2$. Despite its computational efficiency, the competitive ratio of the projection-based algorithm is larger, which will affect the theoretical guarantee of OCFT as discussed in Section 4.

3.3 Policy augmentation

Based on the selected hypothesis model, the algorithm calls the augmentation subroutine on the black-box policy π_t via the following robust optimization (line 6):

$$\min_{a \in \mathcal{A},\ \delta_1, \delta_2 \in \mathbb{R}} \quad \lambda \, \|a - \pi_t(s_t)\|_2^2 + \eta_1 |\delta_1|^2 + \eta_2 |\delta_2|^2 \tag{2a}$$

$$\text{s.t.} \quad \underline{T} + k - \delta_1 \leq \theta_t^\top \Phi(s_t, z_t, a) \leq \overline{T} - k + \delta_2 \tag{2b}$$

$$k = W + \epsilon\,. \tag{2c}$$

We denote this procedure as AUG $(\theta_t,\ \pi_t\ ;\ s_t)$ to emphasize that (2) is instantiated with θ_t, π_t and takes s_t as input. The objective function (2a) minimizes the weighted sum of the total deviation from the action suggested by the black-box policy measured in Euclidean norm, and the thermal constraint violation, which is quantified by the slack variables δ_1 and δ_2. OCFT will generate the same control action as suggested by the black-box policy if it satisfies the thermal comfort constraints under the learned HVAC dynamics. The *trust parameter* λ can be adjusted over time to reflect how much the algorithm trusts the accuracy of the transferred policy that is being updated for optimizing energy consumption, relative to the accuracy of the learned dynamics model in meeting thermal comfort constraints. Meanwhile, the weights $\eta_1,\ \eta_2$ can be chosen based on the operating season to either tighten the upper or lower thermal comfort constraints.

We denote this procedure as AUG $(s_t\ ;\ \theta_t,\ \pi_t)$ to emphasize that (2) is instantiated with θ_t, π_t and takes s_t as input. The objective function (2a) minimizes the weighted sum of the total deviation from the action suggested by the black-box policy measured in Euclidean norm, and the predicted thermal constraint violation, which is quantified by the slack variables δ_1 and δ_2. OCFT will generate the same control action as suggested by the black-box policy if it satisfies the thermal comfort constraints under the learned HVAC dynamics. The trust parameter λ can be adjusted over time to reflect how much the algorithm trusts the accuracy of the transferred policy that is being updated for optimizing energy consumption, relative to the accuracy of the learned dynamics model in meeting thermal comfort constraints. Meanwhile, the weights $\eta_1,\ \eta_2$ can be chosen based on the operating season to either tighten the upper or lower thermal comfort constraints. Additional penalty terms, such as temperature set point tracking, can also be imposed in (2a). In our experiments, the selection of λ significantly influences the performance of the algorithm (Figure 6). It will be important to study how to adjust λ online for Pareto optimality between energy consumption and comfort adherence, which will be future work.

Given the currently selected model parameter θ_t, constraint (2b) ensures that the action generated by (2) will satisfy the thermal comfort constraints with the robustness margin k chosen to limit the effects of disturbances and uncertainties, *if θ_t were the true parameter.* The slack variables δ_1 and δ_2 handle the cases where (2) may become infeasible due to violation of Assumption 3 during deployment. This can happen when the control action available to Algorithm 1 does not have enough control authority to achieve the prescribed thermal comfort levels. For example, if Algorithm 1 controls the position of the damper in a variable air volume box that has a reheat mechanism, then the best that can be done to cool down the corresponding zone is to completely open the damper and turn off the reheat mechanism. But even this action may not be enough to meet the thermal comfort requirement, leading to an infeasible problem without the slack variables.

3.4 Policy update

OCFT is capable of augmenting general black-box transfer learning algorithms. Since many transfer learning methods perform online adaptation at regular intervals to update the policy, OCFT can use the updated policy in the next iteration (line 7). For example, [16] proposed fine-tuning of the transferred policy with data observed online after a certain number of steps. On the other hand, [28] proposed offline policy transfer that only needs a small amount of log data from the target building to select the best policies among a set of diverse and high-quality policies. This policy may be further updated using online data from the target building although it is not essential.

4 THEORETICAL DESIGN INSIGHTS

OCFT is inspired by recent theoretical advances in online control for unknown dynamical systems, where controllers learn to generate actions with safety and performance guarantees by adapting system models online based on the data collected through interaction with the unknown system [18, 24, 26, 31, 32, 41, 55]. This line of work has the common assumption that the underlying system dynamics is either *linear* or *contractive*. However, HVAC zones are known to have nonlinear thermal dynamics, where model learning is challenging due to operational and comfort constraints, and complex energy consumption models. To provide a rigorous foundation on OCFT, we present an analysis in two extreme cases:

a) when there is bounded worst-case model mismatch between (1) and the true HVAC system dynamics, i.e., where such mismatch is modeled by adversarial d_t under Assumption 2

b) when the dynamics model (1) perfectly characterizes the true HVAC system dynamics and the only disturbances in the system are ambient noise, i.e., where d_t is stochastic, i.i.d. zero-mean.

In particular, we show that OCFT simultaneously guarantees finite constraint violation under a) and complete constraint satisfaction in finite time under b).

Theorem 1. *Under Assumptions 1 to 3, Algorithm 1 guarantees that for all $t \geq 0$, the comfort constraints will be violated at most*

$$\frac{2\gamma(q)}{\epsilon} T_{\max} diam(\mathcal{P}_0) + 1 \qquad (3)$$

times for each zone, where q is the number of feature functions, $\gamma(q)$ is the competitive ratio of the chosen NCBC algorithm, $T_{\max}$ is defined in Assumption 1, and $diam(\mathcal{P}_0)$ is defined as $\max_{\theta,\, \theta' \in \mathcal{P}_0} \|\theta - \theta'\|_2$, denoting the diameter of $\mathcal{P}_0$,.

The finite violation bound (3) holds despite potential disturbances that model a variety of realistic uncertainties, such as discretization errors as a result of sampling-based digital control and model mismatch between (1) and the ground truth underlying HVAC system. As shown in Section 5.2, such a general disturbance model is crucial for adapting the same feature functions to a variety of unseen environments. The algorithmic and environmental factors that affect Theorem 1 are:

- Choice of NCBC: As discussed in Section 3.2, different nested CBC algorithms offer distinct trade-offs between competitive ratio $\gamma(q)$ as a function of the dimension of the parameter and computational efficiency. Larger $\gamma(q)$ results in worse dependence on the number of feature functions in (3). For example, the Steiner point achieves $\gamma(q) = q/2$ [10] whereas the projection-based method implemented in the experiments in Section 5.2 has $\gamma(q) = O\left((q-1)q^{q/2}\right)$ [4].
- Robustness margin ϵ: In the context of Assumptions 2 and 3, the margin ϵ will be small if the disturbance bound W is large relative to the thermal comfort levels $\underline{T}$, $\overline{T}$. This implies that Theorem 1 depends on the inherent difficulty of the HVAC control problem even if the true parameter $\theta^\star$ for (1) is known.
- Feature functions Φ: The gradient of (1) with respect to θ is $\Phi(s_t, z_t, a_t)$. Therefore, $T_{\max}$ bounds the sensitivity of the zone temperature prediction according to (1) with respect to θ. Larger $T_{\max}$ means that (1) is more sensitive to model parameter variation, which can lead to more constraint violation when OCFT changes the hypothesis model parameters over time.
- Quality of the log data $\mathcal{D}$: Since OCFT uses log data to generate the initial consistent model set $\mathcal{P}_0$ whose size proportionally affects (3), the more informative $\mathcal{D}$ is in terms of reducing model, the less OCFT will violate comfort constraints online.

The key ingredients for Theorem 1 is the combination of NCBC, which guarantees that the total movement of the selected hypothesis model by Algorithm 1 is at most $\text{diam}(\mathcal{P}_0)$, and AUG, which satisfies Assumption 3 and generates actions that will robustly satisfy the constraints if the selected hypothesis parameter were the true model parameter. We provide a proof in Appendix A, which generalizes the analysis of similar robust online control algorithms in linear systems.

On the other hand, if (1) models the ground truth HVAC dynamics accurately, where disturbances are stochastic noise, OCFT provides a stronger constraint satisfaction guarantee as follows:

Corollary 1 (informal). *If in addition to Assumptions 1 to 3, the disturbances d_t are zero-mean and independently and identically distributed, with certain persistent excitation condition, then with high probability, Algorithm 1 will not violate the thermal comfort constraints for all $t \geq \tilde{O}(q^{2.5}/\epsilon)$, where q is the dimension of the feature functions, and ϵ is the robustness margin from Assumption 3, and $\tilde{O}$ omits logrithmic factors.*

We provide the formal statement of Corollary 1 and proof in Appendix B. The key technical insight that enables Corollary 1 is that in the idealized stochastic setting, the set $\mathcal{P}_t$ will converge to $\theta^\star$ swiftly such that there exists a finite time $T^\star$ where for all $t \geq T^\star$ and for all $\theta \in \mathcal{P}_t$, we have $\left\|\theta - \theta^\star\right\| \leq \epsilon$. Therefore, under Assumption 3, we conclude that for all $t \geq T^\star$, OCFT guarantees thermal constraint satisfaction.

Note that thanks to the additional stochasticity in disturbances and idealized modeling, the total number of constraint violation is polynomial in the number of feature functions regardless of the choice of nested CBC algorithms for NCBC in Corollary 1. This is in contrast with Theorem 1, where the dependency can be exponential.

5 CASE STUDY

To showcase the effectiveness of OCFT and explore its key properties, we consider a case study in EnergyPlus using one of the prototype building models provided by the U.S. Department of Energy.

5.1 Experimental Setup

We consider a three-floor $4{,}982m^2$ office building as the target building for the experiments. Each floor is divided into five thermal zones and equipped with a centralized Air Handling Unit (AHU). The AHU mixes outdoor and return air, and supplies the conditioned air to all thermal zones on that floor. Each zone contains a Variable Air Volume (VAV) system with a damper and a reheat coil. The damper controls the amount of air entering the zone, while the reheat coil heats the supply air to meet the specific temperature requirements of that zone. Figure 2 shows the AHU design and the floor plan of the target building.

To simulate the building dynamics, we use EnergyPlus [47], with COBS [57] serving as the environment interface. At each time step, we observe the following variables: the average temperature and humidity for each zone, outdoor temperature, and solar radiation. We assume the occupancy data is provided, with occupants arriving at the building at 7 a.m. and leaving by 10 p.m., without exception. The building remains unoccupied on Sundays and statutory holidays. The control action adjusts the minimum damper opening percentage in each zone's VAV system, with values ranging from 0.1 to 1. The lower bound of 0.1 is set to satisfy ASHRAE ventilation requirements, ensuring adequate supply of fresh air.

The experiment is conducted across all 19 climate locations (16 in the U.S. and 3 international locations), representing a wide range of global climates. We focus on two extreme months, January and July, to assess control performance under both heating and cooling

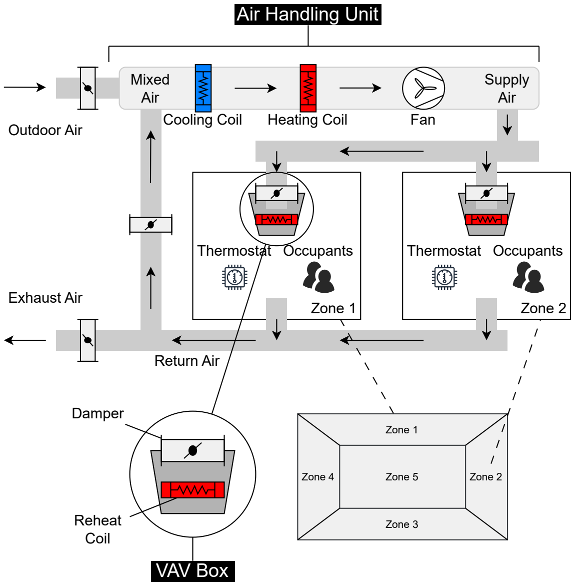

Figure 2: Schematic diagram of the HVAC system in the target building. The three floors have an identical floor plan that is depicted here.

conditions. Climate zones are denoted by a number between 0 and 8 that indicates the temperature profile, and may include a suffix (a, b, or c) to indicate the humidity profile. The control horizon spans 21 days, with 15-minute control time steps.

Prior to the start of the experiment, we collect two weeks of log data when the dampers were controlled by a default controller. This log data is used to warm-start the algorithm and facilitate RL policy transfer as explained later.

We assess thermal comfort satisfaction in each zone by checking whether zone temperature remains between the specified heating and cooling setpoint. We respectively set the heating and cooling setpoints to 20 to 23.5 degrees Celsius in winter, and 23 to 26 degrees Celsius in summer, as recommended by the ASHRAE standard 90.1. Other operational constraints were not included in our experiments, as they can be easily enforced via projection for any black-box policy, including OCFT.

Training and transfer of RL controllers. We used Proximal Policy Optimization (PPO) [42] to learn RL policies through interaction with a small office building in climate zone 5b during January. This building, which serves as our source building, is another DOE prototype building model measuring 511 m^2 and containing only five zones. The heating and cooling setpoints of this building are respectively set to 19 to 25 degrees Celsius in winter, and 22 to 27 degrees Celsius in summer. Note that for both seasons, the acceptable ranges are broader than the respective ranges we considered for the target building. Following [28], we incorporated policy and environmental diversity to obtain 870 zone-level control policies. Using two weeks of log data from our target building (i.e. target building), we then employed the policy selection algorithm proposed in [28] to identify the best policy among the 870 policies for transfer to each zone of that building. This yields 15 policies, each controlling a specific zone in the target building. These black-box policies are augmented using OCFT to ensure constraint satisfaction when deployed in the target building.

ICNNs as feature functions. It is crucial to learn nonlinear feature functions that are convex, as these functions will be embedded in (2). Due to the complexity of building dynamics, linear approximations result in high estimation error. Therefore, we opted to use input convex neural networks (ICNNs) [2] as feature functions, ensuring that the output remains convex with respect to its input. Since we focus on zone-level HVAC control, all ICNNs are trained to model zone-level features. The ICNNs take the current zone and outdoor observations, along with the proposed control action, as input and predict the zone temperature for the next time step. These ICNNs were trained using data from a 15-zone building with a different envelope and HVAC design than the target building, located in climate zone 5b. We trained one ICNN per zone for each month (March, June, September, and December), resulting in 60 ICNN feature functions for zone-level dynamics.

Baseline control methods. We compare OCFT with the following baseline controllers:

(1) **Reinforcement Learning Policy Transfer (RLPT)**: This baseline uses the same policy transfer algorithm that we use in our approach but does not augment the transferred policy using OCFT. As a result, it is not capable of satisfying thermal comfort constraints despite reducing energy consumption.
(2) **Constrained Policy Transfer with Feature Function Adaptation (CPT-Adaptive)**: This baseline uses the same policy transfer algorithm that we use in our approach, yet it has two main differences from our approach. First, it ensures that the policies learned using PPO through interactions with the source building satisfy the source building's constraints by appending a differentiable projection layer [14] to the actor network. This layer maps the agent's proposed action to the closest action that keeps the next-step zone temperature within the acceptable range. To predict the zone temperature in the next time step if a certain action is taken in the current step, we use an ICNN trained for the source building as the feature function. Second, instead of using OCFT to augment the policy selected for transfer, we simply update the ICNN used in the projection layer of that policy and continue using the constrained policy to control the respective zone of the target building, now satisfying the target building's constraints. Specifically, we use the two weeks of log data from the target building to identify the ICNN that predicts the zone temperature most accurately among our pretrained ICNNs. This ICNN will be incorporated in the projection layer of the transferred policy. Compared to our approach, this baseline uses a single feature function in the projection layer, whereas we use multiple feature functions to construct the model that is adapted dynamically.
(3) **Constrained Policy Transfer without Feature Function Adaptation (CPT-NonAdaptive)**: This baseline is similar to CPT-Adaptive with one key difference: the ICNN used in the projection layer of the selected policy is not updated after transfer. Hence, action projection does not guarantee

constraint satisfaction if the transferred ICNN does not approximate dynamics of the target building.

(4) **Default Air System Control Strategy (EnergyPlus)**: This is the default air system control strategy implemented in EnergyPlus, which uses the predictive system energy balance method [47] to predict how much energy must be delivered by HVAC to maintain the desired temperature in each zone.

(5) **Least Squares Estimation (LSE)**: This baseline follows Algorithm 1 but uses least squares estimation in place of NCBC for parameter adaptation.

We note that RLPT, CPT-Adaptive, and CPT-NonAdaptive use the same algorithm for policy selection and transfer, which is the algorithm used in our approach too. They differ in whether and how they enforce constraints after policy transfer. Through a comparison with these baselines, we highlight the effectiveness of OCFT in augmenting a black-box policy that is not learned using a constrained policy optimization algorithm to simultaneously minimize energy consumption and thermal comfort constraint violations.

5.2 Experimental Results

In this section, we present the main result of this paper. We demonstrate that OCFT reduces the total number of thermal comfort constraint violations by 81.28% on average compared to RLPT, the most energy-efficient baseline, across all 19 climate zones during the two extreme seasons (January and July), while consuming 11.23% more energy on average. Additionally, compared to CPT-Adaptive, the policy transfer algorithm with the smallest number of constraint violations among the baselines, OCFT yields 58.70% less thermal constraint violations, with a moderate increase in energy consumption of 8.47%. OCFT surpasses all baselines in constraint satisfaction by integrating all ICNNs and log data, providing a more accurate understanding of system dynamics. In contrast, CPT-Adaptive relies on a single ICNN, which may provide an inaccurate representation of the system, while other baselines fail to account for system dynamics altogether.

We provide further investigation into the properties of the proposed method. In Section 5.2.2, we show that OCFT is capable of trading off energy optimization and comfort constraint adherence via the trust parameter λ. We demonstrate the robustness of OCFT to feature function selection in Section 5.2.3. Moreover, we study the impact of Assumption 2 on the performance of OCFT in Section 5.2.4 where we show that a conservative overestimation of the disturbance bound W is a practical choice for good performance across all 19 climate locations in both extreme seasons.

5.2.1 Constraint Violations. Our first set of experiments focus on constraint violations in the target building, shown in Figure 3. Given the thermal comfort levels, we count one constraint violations whenever a zone temperature is outside of the comfort range. For each method, we tally up the total number of constraint violation over all zones for all 19 climate locations in July and January. Moreover, we fix the nonlinear feature functions and the hyperparameters for OCFT for all experiments for each figure in this section. We refer to Table 1 in the Appendix for a summary of the hyperparameters used for each figure.

As shown in Figure 3 and Figure 4, OCFT achieves the lowest number of thermal constraint violations in nearly all cases, with the exception of climate zones 6b, 7, and 8 during the winter, and climate zone 0b during the summer. In these cases, it performs worse than the EnergyPlus controller, causing 115.2 additional violations per zone on average. These seasons in these climate zones represent extreme weather conditions, such as Extremely Hot Dry (0b) or Arctic/Very Cold (6b, 7, 8), which are either absent or uncommon in the US. As a result, the control policies and feature functions trained on US-based buildings may not generalize well to these extreme weather conditions. Nevertheless, across all climate zones, OCFT significantly outperforms other baselines in terms of reducing constraint violations. On average, it violates thermal comfort constraints only 3.25% of the time (1.64% in summer and 4.87% in winter), equivalent to 131.11 violations per zone over 21 days.

Among the baselines that utilize the policy transfer algorithm, CPT-Adaptive performs the best in terms of constraint satisfaction, achieving the violation rate of 7.87%, which is still more than twice

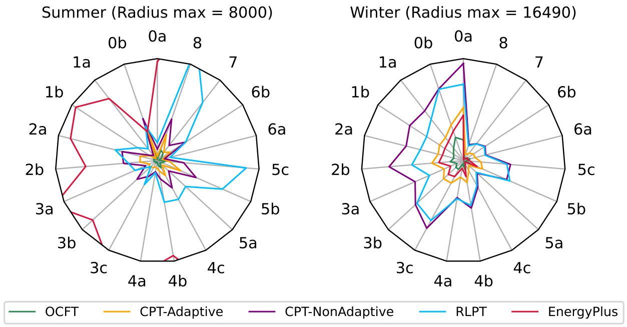

Figure 3: Total number of thermal comfort constraint violation incidents aggregated over all 15 zones for OCFT (green), CPT-Adaptive (orange), CPT-NonAdaptive (purple), RLPT (blue), and default EnergyPlus controller (red) for all 19 climate locations in both summer (left) and winter (right). The closer to the center, the fewer constraint violations. For improved readability, the radius of the radar plot for the summer season is truncated to 8,000 violations so the difference between algorithms becomes clearer.

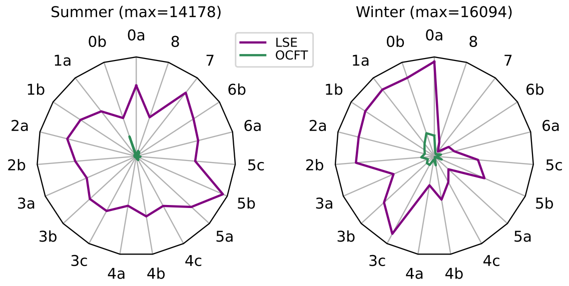

Figure 4: Total number of thermal comfort constraint violation incidents aggregated over all 15 zones for OCFT (green) and LSE (purple) for all 19 climate locations in both summer (left) and winter (right). The closer to the center, the fewer constraint violations.

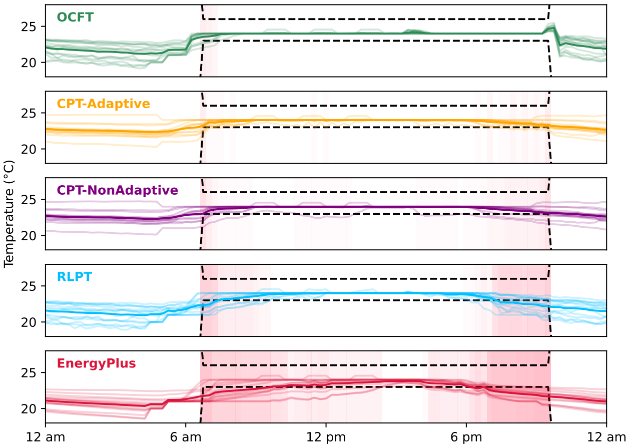

Figure 5: Control behavior of different algorithms over a single day in climate 5b in July. Light lines represent the actual temperature evolution of each zone over time, while the solid line shows the average temperature across all zones. The black dashed lines indicate the comfort temperature bounds, with the pink shaded area representing periods of constraint violations. The darker the shading, the more zones violate the constraint simultaneously.

the violation rate of OCFT. CPT-NonAdaptive, RLPT, and EnergyPlus have higher violation rates of 17.03%, 17.36%, and 19.24%, respectively. The performance of CPT-Adaptive in constraint satisfaction is due to its training process, where the agent is explicitly trained to meet thermal constraints, and the updated feature function provides relatively accurate estimates of the feasible action set. Improvement in constraint satisfaction comes at the cost of a slight increase in energy consumption. OCFT uses 11.23% more energy than the most energy-efficient baseline, RLPT, and 8.47% more energy than CPT-Adaptive. RLPT achieves the lowest energy consumption because it ignores temperature constraints during training and policy transfer, allowing the agent to prioritize energy savings over maintaining thermal comfort.

Compared to LSE, OCFT reduces constraint violations by an impressive 95% on average. This is meant to be an ablation study on the importance of NCBC as part of OCFT, since LSE and OCFT only differ by how model features are adapted online. In particular, this result shows that NCBC is indeed significantly more effective at leveraging online data to reduce model uncertainty for constraint satisfaction than the classic approach of least squares estimation, which is a popular model adaptation method in online control.

These results highlight the potential of OCFT to effectively balance reducing thermal constraint violations while maintaining reasonable energy efficiency, demonstrating its suitability for applications requiring strict adherence to thermal control requirements.

We provide a detailed comparison of OCFT, CPT-Adaptive, CPT-NonAdaptive, RLPT, and the EnergyPlus controller in Figure 5, where the zone temperature of all zones and the averaged value are plotted. The figure illustrates the control behavior of these algorithms during a one-day experiment in climate 5b in July. Light lines represent the temperature changes for individual zones, and the black dashed lines mark the upper and lower bounds of the preferred temperature range, based on thermal comfort constraints. An effective control algorithm should keep the temperature within these bounds during occupied hours. For unoccupied hours, we set the bounds to 13 and 35 degrees Celsius, as per ASHRAE guidelines, though these are not shown in the figure due to the y-axis scale. The shaded areas indicate periods when temperature violations occur, with darker shading representing a higher number of simultaneous violations. This visualization highlights how each algorithm manages temperatures and handles constraint violations throughout the day. Notably, OCFT minimizes the number of violations by learning to preheat the building before occupancy, as well as extending the conditioning period after the building is unoccupied to reduce violations when the preferred temperature bounds shift.

Another interesting observation is the end-of-day behavior. All baselines take actions that gradually reduce zone temperature, relying on the gap between the previous stable temperature and the lower constraint bound to meet thermal comfort requirements in the last few hours of the occupancy period. This approach risks some zones barely remaining above the lower limit by the end of the occupancy period. In contrast, OCFT consumes a little more energy to raise the temperature in late afternoon, ensuring compliance with thermal constraints.

5.2.2 *Navigating energy-comfort trade-off via varying λ.*

The trade-off between optimizing energy consumption and satisfying thermal comfort has been widely observed in HVAC control. As discussed in Section 5.2.1, OCFT reduces thermal comfort violations but at the cost of increased energy usage. In this section, we explore the primary hyperparameter, λ, which governs the energy-comfort trade-off in OCFT. We fix the regularization factors η_1 and η_2 to both be 1, and test λ values ranging from $\{1e1, 5e1, 1e2, 1e3, 2e3, 5e3, 1e4, 1e5, 1e6\}$ across all climate zones and seasons. Since the actions are limited to the range $[0.1,\ 1]$ and the constraint violations are regularly observed to be in $[0,\ 50]$, we choose the regularization penalty in the proposed range to cover a wide range of values to test the balance between selecting actions that are close to the suggested actions and actions that violate thermal comfort constraints. A higher λ makes OCFT more inclined to follow RLPT actions for energy efficiency, with $\lambda = 1e6$ producing behavior nearly identical to RLPT. To enhance readability, Figure 6 presents the percentage increase in energy consumption on the x-axis and the percentage decrease in temperature violations on the y-axis (reversed). Each line connects results from the same climate and season, while the orange star indicates the average change in energy consumption and thermal comfort violations for each λ.

In the winter, a clear trade-off curve emerges across all climates. As the points are connected in order of decreasing λ, the plot demonstrates how λ can effectively tune the balance between energy consumption and temperature constraint satisfaction. Lower λ values result in higher energy consumption but significantly fewer temperature violations. The consistent curve shapes across climates show that OCFT performs robustly in different environments.

In the summer, energy consumption remains relatively stable no matter whether the model is tuned for either stricter temperature constraints (low λ) or more aggressive energy savings (high λ). This suggests that OCFT can substantially reduce constraint

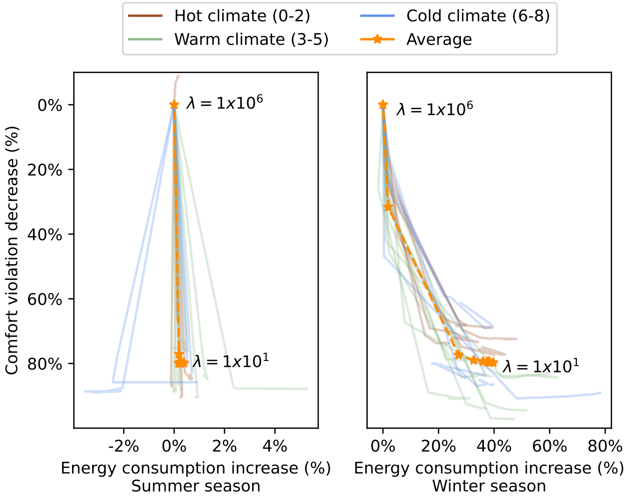

Figure 6: Energy-comfort trade-off of OCFT as the trust parameter λ varies. This parameter balances between following RLPT actions for energy efficiency and selecting actions to minimize comfort constraint violations. Each line represents the change in energy consumption and violations compared to $\lambda = 10^6$ for a specific climate zone and season. The dashed line indicates the average performance across all climates for the given season.

violations without a corresponding increase in energy usage when using a smaller λ, highlighting its potential for efficient operation. Interestingly, in colder climates (zones 6-8), OCFT can reduce both constraint violations and energy consumption simultaneously in the summer, suggesting that a smaller λ may be particularly advantageous in these regions.

5.2.3 *Robustness to feature functions* Φ .

An essential component of OCFT is the set of feature functions, trained on the source building using offline data. While high-quality and diverse feature functions are desirable to better approximate the true HVAC dynamics as described by (1), training a large set of rich feature functions can be resource-intensive. To address this, we investigate the robustness of OCFT relative to both the quality and quantity of feature functions. Specifically, we compare the constraint satisfaction performance of OCFT when using 60 ICNNs, versus reduced versions with only 45 and 30 ICNNs.

The results are summarized in Figure 7. Across all 19 climate locations, OCFT with 30 or 45 ICNNs, randomly sampled from the original set of 60 ICNNs, performs almost as well as the version using all 60 ICNNs. Such results are statistically significant, as indicated by the short error bars. Since we trained 15 ICNNs for each seasonal month, random sampling from the full set of 60 ICNNs can result in an imbalanced representation of the seasonal HVAC dynamics.

Interestingly, in some cases, OCFT performs better with fewer ICNNs. Notably, in climates 2a and 2b, using 30 or 45 ICNNs leads to approximately 10% fewer temperature constraint violations compared to the version using all 60 ICNNs. One possible explanation is that reducing the number of ICNNs could simplify the optimization process during policy learning. In scenarios where HVAC dynamics are not complex, having fewer ICNNs might prevent the model from overfitting.

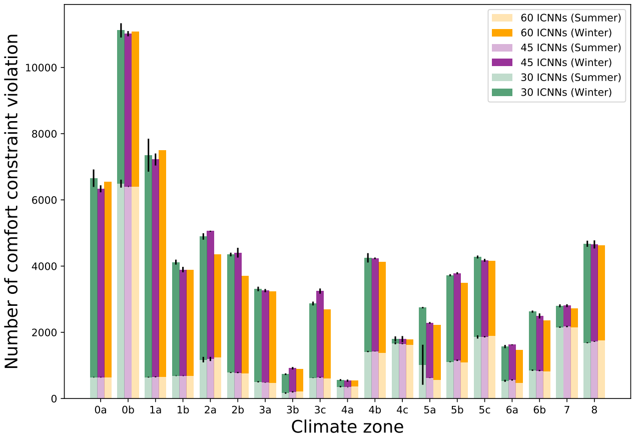

Figure 7: Total number of thermal constraint violation of the 21-day period aggregated over all 15 zones for OCFT with 60 ICNN feature functions (orange), 45 ICNNs uniformly randomly sampled from the original 60 feature functions (purple), and 30 ICNNs randomly sampled from the original 60 feature functions (green) for all 19 climate zones in both summer (lighter shade) and winter (darker shade). Black error bar denotes 1 standard deviation across 3 different seeds.

In summary, Figure 7 suggests that OCFT maintains robust performance even with a significantly reduced number of feature functions. The experimental result demonstrates that reducing the complexity of the feature set does not necessarily compromise, and may even enhance, constraint satisfaction in certain climates.

5.2.4 *Sensitivity to disturbance bound* W.

OCFT requires the knowledge of the upper bound W on the disturbances per Assumption 2. However, in realistic scenarios it is challenging to obtain such an exact bound, especially with feature function approximation (1) for the dynamics. Therefore, one may need to estimate W in order to deploy OCFT. This section investigates the effect of the over- or underestimation of W.

Underestimation of W. For a fixed set of data, disturbance bound W directly controls the size of the consistent model set $\mathcal{P}_t$ (c.f. line 4 of Algorithm 1). If W underestimates the true disturbance bound, eventually $\mathcal{P}_t$ will become empty. Empty $\mathcal{P}_t$ means that there does not exist a parameter in the entire uncertainty set Θ that could have generated the observed data assuming the disturbances are bounded by W. Therefore, we must adjust W to be a higher value and re-compute $\mathcal{P}_t$ such that $\mathcal{P}_t$ is non-empty and continue running OCFT with the new W. This resets the model parameter learning process via NCBC, since previously invalidated parameters under the smaller bound are now re-introduced back to $\mathcal{P}_t$. As a result, constraint violations will continue to happen until W converges upward to the true disturbance bound.

Overestimation of W. If W overestimates the true bound, then $\mathcal{P}_t$ will remain non-empty. However, it has been shown that even

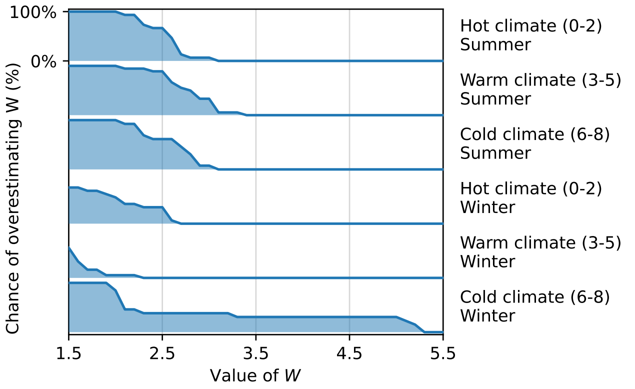

Figure 8: Sensitivity analysis for the value of W. The experiment begins with $W = 1.5$ and tracks the likelihood of the model set $\mathcal{P}_t$ becoming empty throughout the experiment. W is increased by 0.1 each time $\mathcal{P}_t$ is empty.

in the idealized stochastic disturbance and linear dynamics setting, if W is conservative, the set $\mathcal{P}_t$ will not converge to the true parameter $\theta^\star$ but only to a neighborhood of it, with the size of the neighborhood depending on the conservativeness of W [31]. Moreover, unnecessarily large W will generate a large initial consistent model set $\mathcal{P}_0$, which negatively impacts the constraint violation bound in (3).

Identifying the optimal balance between over- and underestimating W can be challenging. Different environments and HVAC systems may necessitate distinct values of W for the algorithm to perform optimally. To address this, we examine the sensitivity of OCFT's performance with respect to variations in W. Figure 8 illustrates the probability that a given W results in an empty $\mathcal{P}_t$ when running experiments across various climates and seasons. A higher percentage indicates that the corresponding W is likely underestimated for the specific experiment. It is evident that for $W > 3.5$, the risk of underestimation is minimal, except in cold winter conditions. This figure highlights the minimum W needed to avoid an empty $\mathcal{P}_t$ during experiments. Furthermore, Figure 8 shows that, regardless of climate zone, setting W between 3.5 and 5.5 is sufficient in most cases in our study. This indicates that extensive tuning of W is generally unnecessary across different settings. In all the experiments presented in the previous sections, we fix $W = 4$ for Algorithm 1 and increase this bound by $\Delta W = 0.1$ every time $\mathcal{P}_t$ becomes empty.

In Figure 9, we show the performance of OCFT instantiated with different values of W corresponding to that of Figure 8. For clarity, the result is grouped based on the climate zone. Moreover, we calculate the average number of constraint violation over each group. As can be seen in Figure 9, even with a highly conservative overestimation of W, e.g., $W = 6$, OCFT still manages to achieve similar constraint satisfaction performance as OCFT with $W = 4$, where Figure 8 has shown this value is sufficient for feasibility for hot and warm climate zones. Therefore, the experiment suggests that despite theoretical drawbacks, OCFT using a conservative upper bound for W can potentially remain performant for HVAC control applications across diverse environments.

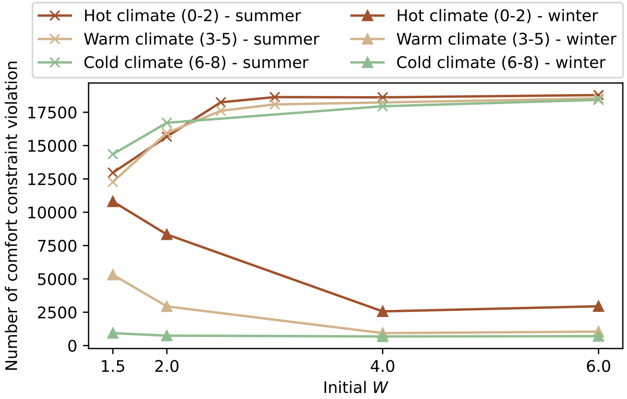

Figure 9: Average number of temperature constraint violations observed in experiments with varying W values across different climates and seasons.

6 CONCLUDING REMARKS

Constraint satisfaction is a central challenge in machine learning-based control of HVAC systems due to the inherent tension between exploration and safety. Despite the growing body of work on optimal control of HVAC systems, limited progress has been made toward enforcing operational and thermal comfort constraints, especially by augmenting existing control policies that are learned on the target building or transferred to that building, rather than developing new control policies which is prohibitively expensive. We addressed this important gap in the literature by designing a novel algorithm, called OCFT, and experimentally quantified the cost of compliance with constraints with respect to energy consumption. Using learned feature functions that approximate zone-level dynamics in a source building, OCFT augments an arbitrary black-box policy by solving a robust optimization problem to select actions that closely track actions suggested by the black-box policy, while conforming to operational and thermal comfort constraints. This results in a significant reduction of constraint violations across a wide range of climates. An important future direction is to study the theoretical properties of the trust parameter in the algorithm used to trade off constraint satisfaction and energy consumption.

REFERENCES

[1] Hüseyin Akçay. 2004. The size of the membership-set in a probabilistic framework. *Automatica* 40, 2 (2004), 253–260.

[2] Brandon Amos, Lei Xu, and J. Zico Kolter. 2017. Input Convex Neural Networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 146–155.

[3] Zhiyu An, Xianzhong Ding, Arya Rathee, and Wan Du. 2023. CLUE: Safe Model-Based RL HVAC Control Using Epistemic Uncertainty Estimation. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. ACM, 149–158.

[4] CJ Argue, Sébastien Bubeck, Michael B Cohen, Anupam Gupta, and Yin Tat Lee. 2019. A nearly-linear bound for chasing nested convex bodies. In *Proc. 30th Annu. ACM-SIAM Symp. Discrete Algorithms*. 117–122.

[5] CJ Argue, Anupam Gupta, Ziye Tang, and Guru Guruganesh. 2021. Chasing convex bodies with linear competitive ratio. *Journal of the ACM (JACM)* 68, 5 (2021), 1–10.

[6] Javier Arroyo, Carlo Manna, Fred Spiessens, and Lieve Helsen. 2022. Reinforced model predictive control (RL-MPC) for building energy management. *Applied Energy* 309 (2022), 118346.

[7] Anil Aswani, Humberto Gonzalez, S Shankar Sastry, and Claire Tomlin. 2013. Provably safe and robust learning-based model predictive control. *Automatica* 49, 5 (2013), 1216–1226.

[8] Er-Wei Bai, Roberto Tempo, and Hyonyong Cho. 1995. Membership set estimators: size, optimal inputs, complexity and relations with least squares. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* 42, 5 (1995), 266–277.

[9] Nibodh Boddupalli, Aqib Hasnain, Sai Pushpak Nandanoori, and Enoch Yeung. 2019. Koopman operators for generalized persistence of excitation conditions for nonlinear systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 8106–8111.

[10] Sébastien Bubeck, Bo'az Klartag, Yin Tat Lee, Yuanzhi Li, and Mark Sellke. 2020. Chasing nested convex bodies nearly optimally. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (Salt Lake City, Utah). SIAM, SIAM, USA, 1496–1508. https://doi.org/10.1137/1.9781611975994.91

[11] Sébastien Bubeck, Yin Tat Lee, Yuanzhi Li, and Mark Sellke. 2019. Competitively chasing convex bodies. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 861–868.

[12] Kathleen Champion, Bethany Lusch, J Nathan Kutz, and Steven L Brunton. 2019. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences* 116, 45 (2019), 22445–22451.

[13] Bingqing Chen, Zicheng Cai, and Mario Bergés. 2019. Gnu-RL: A Precocial Reinforcement Learning Solution for Building HVAC Control Using a Differentiable MPC Policy. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. ACM, 316–325.

[14] Bingqing Chen, Priya L. Donti, Kyri Baker, J. Zico Kolter, and Mario Bergés. 2021. Enforcing Policy Feasibility Constraints through Differentiable Projection for Energy Optimization. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*. ACM, 199–210.

[15] Yize Chen, Yuanyuan Shi, and Baosen Zhang. 2019. Optimal control via neural networks: A convex approach, In Proceedings of the Seventh International Conference on Learning Representations. *arXiv preprint arXiv:1805.11835*.

[16] Davide Coraci, Silvio Brandi, Tianzhen Hong, and Alfonso Capozzoli. 2023. Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. *Applied Energy* 333 (2023), 120598.

[17] Iago Cupeiro Figueroa, Ján Drgoňa, and Lieve Helsen. 2019. State estimators applied to a white-box geothermal borefield controller model. In *Building Simulation 2019*, Vol. 16. IBPSA, 830–837.

[18] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. 2019. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics* (2019), 1–47.

[19] Xianzhong Ding, Wan Du, and Alberto E. Cerpa. 2020. MB2C: Model-Based Deep Reinforcement Learning for Multi-zone Building Control. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. ACM, 50–59.

[20] Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, Juraj Oravec, Michael Wetter, Draguna L. Vrabie, and Lieve Helsen. 2020. All you need to know about model predictive control for buildings. *Annual Reviews in Control* 50 (2020), 190–232.

[21] Xi Fang, Guangcai Gong, Guannan Li, Liang Chun, Pei Peng, Wenqiang Li, and Xing Shi. 2023. Cross temporal-spatial transferability investigation of deep reinforcement learning control strategy in the building HVAC system level. *Energy* 263 (2023), 125679.

[22] Zeyu Feng, Bowen Zhang, Jianxin Bi, and Harold Soh. 2023. Safety-constrained policy transfer with successor features. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7219–7225.

[23] Dimitry Gorinevsky. 1995. On the persistency of excitation in radial basis function network identification of nonlinear systems. *IEEE Transactions on Neural Networks* 6, 5 (1995), 1237–1244.

[24] Elad Hazan, Sham Kakade, and Karan Singh. 2020. The nonstochastic control problem. In *Algorithmic Learning Theory*. PMLR, 408–421.

[25] Dimitar Ho. 2020. A system level approach to discrete-time nonlinear systems, In 2020 American Control Conference (ACC). *ACC*, 1625–1630.

[26] Dimitar Ho, Hoang Le, John Doyle, and Yisong Yue. 2021. Online Robust Control of Nonlinear Systems with Large Uncertainty, In Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS) (San Diego, CA, USA). *arXiv preprint arXiv:2103.11055* 130, 3475–3483.

[27] Yangsheng Hu, Li Tan, and Raymond A de Callafon. 2019. Persistent excitation condition for MIMO Volterra system identification with Gaussian distributed input signals. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 1752–1757.

[28] Aakash Krishna G.S., Tianyu Zhang, Omid Ardakanian, and Matthew E. Taylor. 2023. Mitigating an adoption barrier of reinforcement learning-based control strategies in buildings. *Energy and Buildings* 285 (2023), 112878.

[29] Li Lan, Pawel Wargocki, and Zhiwei Lian. 2011. Quantitative measurement of productivity loss due to thermal discomfort. *Energy and Buildings* 43, 5 (2011), 1057–1062.

[30] Antoine Leeman, Johannes Köhler, Samir Bennani, and Melanie Zeilinger. 2023. Predictive safety filter using system level synthesis. In *Learning for Dynamics and Control Conference*. PMLR, 1180–1192.

[31] Yingying Li*, Jing Yu*, Lauren Conger, Taylan Kargin, and Adam Wierman. 2024. Learning the Uncertainty Sets of Linear Control Systems via Set Membership: A Non-asymptotic Analysis. *Forty-first International Conference on Machine Learning (ICML)* (2024). https://openreview.net/forum?id=n2kq2EOHFE

[32] Yiheng Lin, James A Preiss, Fengze Xie, Emile Anand, Soon-Jo Chung, Yisong Yue, and Adam Wierman. 2024. Online Policy Optimization in Unknown Nonlinear Systems. *arXiv preprint arXiv:2404.13009* (2024).

[33] Hsin-Yu Liu, Bharathan Balaji, Sicun Gao, Rajesh Gupta, and Dezhi Hong. 2022. Safe HVAC Control via Batch Reinforcement Learning. In *Proceedings of the ACM/IEEE 13th International Conference on Cyber-Physical Systems*. 181–192.

[34] Hsin-Yu Liu, Bharathan Balaji, Rajesh Gupta, and Dezhi Hong. 2023. Rule-based Policy Regularization for Reinforcement Learning-based Building Control. In *Proceedings of the 14th ACM International Conference on Future Energy Systems*. ACM, 242–265.

[35] Xiaonan Lu, Mark Cannon, and Denis Koksal-Rivet. 2019. Robust adaptive model predictive control: Performance and parameter estimation. *International Journal of Robust and Nonlinear Control* (2019).

[36] Yudong Ma, Garrett Anderson, and Francesco Borrelli. 2011. A distributed predictive control approach to building temperature regulation. In *Proceedings of the 2011 American Control Conference*. IEEE, 2089–2094.

[37] Alexandre Mauroy, Y Susuki, and Igor Mezic. 2020. *Koopman operator in systems and control*. Springer.

[38] Srinarayana Nagarathinam, Vishnu Menon, Arunchandar Vasan, and Anand Sivasubramaniam. 2020. MARCO - Multi-Agent Reinforcement learning based COntrol of building HVAC systems. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*. ACM, 57—-67.

[39] Kumpati S Narendra and Anuradha M Annaswamy. 1987. Persistent excitation in adaptive systems. *Internat. J. Control* 45, 1 (1987), 127–160.

[40] Kingsley Nweye, Siva Sankaranarayanan, and Zoltan Nagy. 2023. MERLIN: Multi-agent offline and transfer learning for occupant-centric operation of grid-interactive communities. *Applied Energy* 346 (2023), 121323.

[41] Venkatraman Renganathan, Andrea Iannelli, and Anders Rantzer. 2023. An online learning analysis of minimax adaptive control. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 1034–1039.

[42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. preprint (2017), 9 pages. arXiv:1707.06347

[43] Mark Sellke. 2020. Chasing convex bodies optimally. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1509–1518.

[44] Gianluca Serale, Massimo Fiorentini, Alfonso Capozzoli, Daniele Bernardini, and Alberto Bemporad. 2018. Model predictive control (MPC) for enhancing building and HVAC system energy efficiency: Problem formulation, applications and opportunities. *Energies* 11, 3 (2018), 631.

[45] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. 2018. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*. PMLR, 439–473.

[46] Matthew E Taylor and Peter Stone. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10, 7 (2009), 1633–1685.

[47] U.S. Department of Energy. 2024. EnergyPlus Version 24.1.0 Documentation. https://energyplus.net/assets/nrel_custom/pdfs/pdfs_v24.1.0/EngineeringReference.pdf.

[48] Kim P Wabersich and Melanie N Zeilinger. 2018. Linear model predictive safety certification for learning-based control. In *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 7130–7135.

[49] Zhe Wang and Tianzhen Hong. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* 269 (2020), 115036.

[50] Haonan Xu and Yingying Li. 2024. On the Convergence Rates of Set Membership Estimation of Linear Systems with Disturbances Bounded by General Convex Sets. *arXiv preprint arXiv:2406.00574* (2024).

[51] Shichao Xu, Yixuan Wang, Yanzhi Wang, Zheng O'Neill, and Qi Zhu. 2020. One for Many: Transfer Learning for Building HVAC Control. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. ACM, 230–239.

[52] Qisong Yang, T Simão, Nils Jansen, S Tindemans, and M Spaan. 2022. Training and transferring safe policies in reinforcement learning. In *Proceedings of the Adaptive and Learning Agents Workshop*. AAMAS.

[53] Yihang Yao, Zuxin Liu, Zhepeng Cen, Jiacheng Zhu, Wenhao Yu, Tingnan Zhang, and Ding Zhao. 2024. Constraint-conditioned policy optimization for versatile safe reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

[54] Christopher Yeh, Jing Yu, Yuanyuan Shi, and Adam Wierman. 2024. Online learning for robust voltage control under uncertain grid topology. *IEEE Transactions on Smart Grid* (2024).

[55] Jing Yu, Dimitar Ho, and Adam Wierman. 2023. Online Adversarial Stabilization of Unknown Networked Systems. In *Prof. ACM Meas. Anal. Comput. Syst.*, Vol. 7. 1–43.

[56] Chi Zhang, Sanmukh Rao Kuppannagari, and Viktor K. Prasanna. 2022. Safe Building HVAC Control via Batch Reinforcement Learning. *IEEE Transactions on Sustainable Computing* 7, 4 (2022), 923–934.

[57] Tianyu Zhang and Omid Ardakanian. 2020. COBS: COmprehensive Building Simulator. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation.* ACM, 314–315.

[58] Tianyu Zhang, Gaby Baasch, Omid Ardakanian, and Ralph Evins. 2021. On the Joint Control of Multiple Building Systems with Reinforcement Learning. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems.* ACM, 60–72.

A PROOF OF THEOREM 1

The main technical machinery that enables Theorem 1 is the following result adapted in the context of this paper, which provides a sufficient condition on (2) for Theorem 1 to hold.

Corollary 2 (Direct consequence of Theorem 2.5 in [26]). *If there exists $\rho > 0$ such that for all $\theta' \in \mathcal{P}_0$ and $\hat{\theta} \in \mathcal{P}_0$ such that $\left\|\theta' - \hat{\theta}\right\|_2 \le \rho$, the action a' generated by the solution of (2) using θ' guarantees that $\underline{T} + W \le \hat{\theta}^\top \Phi(s, z, a') \le \overline{T}$-$W$ for all $s \in \mathcal{S}, z \in \mathcal{Z}$, then the following violation guarantees hold: Algorithm 1 will violate the constraint bound $[\underline{T}, \overline{T}]$ at most*

$$\frac{2\gamma}{\rho} diam(\mathcal{P}_0) + 1$$

times, where γ denotes the competitive ratio of NCBC, and $\mathcal{P}_0$ is the initial consistent parameter set constructed using log data.

We refer interested readers to [26] for the details of the original result. With Corollary 2, the proof of Theorem 1 reduces to showing that indeed there exists ρ for (2).

We show that $\rho = \epsilon / T_{\max}$ for (2) satisfy the requirement of Corollary 2. To see this, suppose action a has been generated by (2) using θ based on the current states s and z. Consider $\hat{\theta} \in \mathcal{P}_0$ such that $\left\|\hat{\theta} - \theta\right\|_2 \le \rho$, then we know by Assumption 3 that

$$\underline{T} + W + \epsilon \le \theta^\top \Phi(s, z, a) \le \overline{T} - W - \epsilon.$$

Therefore,

$$\underline{T} + W + \epsilon \le (\theta - \hat{\theta})^\top \Phi(s, z, a) + \hat{\theta}^\top \phi(s, z, a) \le \overline{T} - W - \epsilon.$$

and with the choice of $\rho = \epsilon / T_{\max}$ we have

$$\begin{aligned}
&(\theta - \hat{\theta})^\top \Phi(s, z, a) + \hat{\theta}^\top \phi(s, z, a) \\
&\le \|\theta - \hat{\theta}\|_2 \|\Phi(s, z, a)\|_2 + \hat{\theta}^\top \phi(s, z, a) \\
&\le \rho T_{\max} + \hat{\theta}^\top \phi(s, z, a).
\end{aligned}$$

So $\hat{\theta}^\top \phi(s, z, a) \le \overline{T} - W$. The lower bound can be shown in a symmetric fashion, which shows the desired result.

B PROOF OF COROLLARY 1

Before we present the formal state and proof of Corollary 1, we will first introduce the technical assumption and definition necessary for the result. Corollary 1 considers the idealized setting where the underlying HVAC system can be described by (1) , with d_t modeled as stochastic noise. To this end, we formally define the additional assumption on d_t.

Assumption 4 (Tight bound on d_t). *For any $\epsilon > 0$, there exists $\phi_d(\epsilon) > 0$, such that we have*

$$\min(\mathbb{P}(d_t \le \epsilon - W), \mathbb{P}(d_t \ge W - \epsilon)) \ge \phi_d(\epsilon).$$

Moreover, $\phi_d(\epsilon) = \Omega(\epsilon)$.

This assumption essentially says that the bound on the disturbances per Assumption 2 should be tight, where there is nontrivial probability for d_t to take values that are arbitrarily close to the bound W and $-W$. In particular, the probability of being ϵ near the boundaries should scale at least linearly with ϵ. Many common distributions such as the uniform distribution and truncated Gaussian distribution satisfy this requirement.

Assumption 5 (Block Martingale small-ball [45]). *With filtration $\mathcal{F}_t = \mathcal{F}(d_0 \ldots, d_{t-1}, z_0, \ldots, z_{t-1}, a_0, \ldots, a_{t-1}, s_0, \ldots, s_t)$, there exists constants $\sigma > 0$ and $0 < p \le 1$ such that for all $\lambda \in \mathbb{R}^q$ such that $\|\lambda\|_2 = 1$, the $\mathcal{F}_t$-adapted stochastic process $\{\Phi(s_t, z_t, a_t)_t\}_{t \ge 0}$ satisfies $\mathbb{P}(|\lambda^\top \Phi(s_t, z_t, a_t)| \ge \sigma | \mathcal{F}_t) \ge p$ for all $t \ge 1$.*

Assumption 5 requires the actions a_t and measurable disturbances z_t to generate $\Phi(s_t, z_t, a_t)$ that satisfy the block martingale small-ball (BMSB) condition. BMSB can be interpreted as the probabilistic counterpart of persistent excitation, commonly seen in the system identification literature [39]. BMSB essentially states that the state vector $\Phi(s_t, z_t, a_t)$ produced by the feature functions will visit all directions in the state space, facilitating the identification task. Depending on the type of feature functions, Assumption 5 can be guaranteed using a_t based on ideas from e.g.[9, 23, 27]. With Assumption 5, we are in position to state the following theorem, which Corollary 1 builds upon.

Theorem 2 (adapted from Theorem 3.1 of [31]). *Suppose Assumption 1, 2, 4 and 5 holds and d_t is i.i.d. with zero-mean for all $t \ge 0$. For any $m > 0$ any $\epsilon > 0$, when $T > m$, we have*

$$\mathbb{P}(diam(\mathcal{P}_T) > \epsilon) \le \frac{T}{m} \tilde{O}(q^{2.5}) a_2^q \exp(-a_3 m) \tag{4a}$$

$$+ \tilde{O}(q^{2.5}) a_4^q \left(1 - \phi_d\left(\frac{a_1 \epsilon}{4}\right)\right)^{\lceil T/m \rceil} \tag{4b}$$

where $\tilde{O}$ ignores logarithmic terms, q is the number of feature functions in (2), $a_1 = \frac{\sigma p}{4}$, $a_2 = \frac{64(T_{\max}\kappa)^2}{\sigma^2 p^2}$, $a_3 = \frac{p^2}{8}$, $a_4 = \frac{4T_{\max}\kappa\sqrt{q}}{a_1}$, p, σ are defined in Assumption 5, $\lceil \cdot \rceil$ denotes the ceiling function, $\kappa := \max_{\theta \in \Theta} \|\theta\|_2$ with Θ the initial uncertainty set defined in Assumption 1, and $\phi_d(\cdot)$ is defined in Assumption 4.

Theorem 2 guarantees that under the specified stochastic assumptions on d_t and the output of the feature functions, the diameter of the consistent model parameter set $\mathcal{P}_T$ will be arbitrarily small over time with high probability. Since the true model parameter $\theta^\star \in \mathcal{P}_T$ for all $T \ge 0$, any hypothesis models selected by NCBC in Algorithm 1 therefore will be arbitrarily close to $\theta^\star$ under Theorem 2. For the full constants inside $\tilde{O}(\cdot)$, we refer interested readers to [31]. We are now ready to state the formal version of Corollary 1.

Corollary 3 (Sample complexity). *Suppose Assumption 1, 2, 4 and 5 holds and d_t is i.i.d. with zero-mean for all $t \ge 0$. Then with at least probability $1 - 2\delta$ for any $\delta > 0$, Algorithm 1 will not violate the comfort constraints $[\underline{T}, \overline{T}]$ for all $t \ge T^\star$ with $T^\star = \tilde{O}(q^{2.5}/\epsilon)$, where ϵ is the robustness margin in Assumption 3.*

Proof. Fix a small constant $\delta > 0$, our goal is to find $T^\star$ that will make the right hand side of (4) equal to 2δ in order to reach the conclusion of the corollary. To achieve this, we will choose $m = O(q + \log T + \log(1/\epsilon))$. Plugging this m back in (4a), we see that $T\tilde{O}(q^{2.5})a_2^q \exp(-a_3 m) \leq \epsilon$. Since $m \geq 1$, we obtain (4a)$\leq \delta$.

For (4b), we use the assumption that $\phi_d(\frac{a_1\epsilon}{4}) = O(\frac{a_1\epsilon}{4})$ to obtain that $(1 - O(\frac{a_1\epsilon}{4})) = \left(\frac{\delta}{\tilde{O}(q^{2.5})a_4^q}\right)^{m/T}$, which is equivalent to

$$\begin{aligned}\epsilon &= O(\frac{4}{a_1})\left(1 - \left(\frac{\delta}{\tilde{O}(q^{2.5})a_4^q}\right)^{m/T}\right)\\ &\leq O(\frac{4}{a_1})\log\left(\left(\frac{\delta}{\tilde{O}(q^{2.5})a_4^q}\right)^{m/T}\right)\\ &= \tilde{O}\left(\frac{q^{2.5}}{T}\right).\end{aligned}$$

Here we have assumed without loss of generality that T/m is an integer. Therefore, for all $T \geq \tilde{O}(q^{2.5}/\epsilon)$, we have that $\mathrm{diam}(\mathcal{P}_T) \leq \epsilon$ with probability $1 - 2\delta$.

□

C HYPERPARAMETERS FOR EXPERIMENTS

Table 1: Hyperparameters used for each figure.

	Hyperparameter name		
	W	λ	ICNN size
Figure 3, 4, 5	4	50	60
Figure 6	4	9 values	60
Figure 7	4	50	3 values
Figure 8	1.5	50	60
Figure 9	4 values	50	3 values