



PDF Download
3679240.3734595.pdf
25 January 2026
Total Citations: 0
Total Downloads: 701

Latest updates: <https://dl.acm.org/doi/10.1145/3679240.3734595>

RESEARCH-ARTICLE

Learning-Augmented Online Control for Decarbonizing Water Infrastructures

JIANYI YANG, University of Houston, Houston, TX, United States

PENGFEI LI, University of California, Riverside, Riverside, CA, United States

TONGXIN LI, The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong, China

ADAM WIERMAN, California Institute of Technology, Pasadena, CA, United States

SHAOLEI REN, University of California, Riverside, Riverside, CA, United States

Open Access Support provided by:

California Institute of Technology

University of Houston

The Chinese University of Hong Kong, Shenzhen

University of California, Riverside

Published: 17 June 2025

[Citation in BibTeX format](#)

E-Energy '25: The 16th ACM International Conference on Future and Sustainable Energy Systems
June 17 - 20, 2025
Rotterdam, Netherlands

Conference Sponsors:
SIGENERGY

Learning-Augmented Online Control for Decarbonizing Water Infrastructures

Jianyi Yang
University of Houston
Houston, USA
jyang71@central.uh.edu

Pengfei Li
UC Riverside
Riverside, USA
pli081@ucr.edu

Tongxin Li
The Chinese University of Hong
Kong, Shenzhen
Shenzhen, China
litongxin@cuhk.edu.cn

Adam Wierman
Caltech
Pasadena, USA
adamw@caltech.edu

Shaolei Ren
UC Riverside
Riverside, USA
shaolei@ucr.edu

Abstract

Water infrastructures are essential for drinking water supply, irrigation, fire protection, and other critical applications. However, water pumping systems, which are key to transporting water to the point of use, consume significant amounts of energy and emit millions of tons of greenhouse gases annually. With the wide deployment of digital water meters and sensors in these infrastructures, Machine Learning (ML) has the potential to optimize water supply control and reduce greenhouse gas emissions. Nevertheless, the inherent vulnerability of ML methods in terms of worst-case performance raises safety concerns when deployed in critical water infrastructures. To address this challenge, we propose a learning-augmented online control algorithm, termed LAOC, designed to dynamically schedule the activation and/or speed of water pumps. To ensure safety, we introduce a novel design of safe action sets for online control problems. By leveraging these safe action sets, LAOC can provably guarantee safety constraints while utilizing ML predictions to reduce energy and environmental costs. Our analysis reveals the tradeoff between safety requirements and average energy/environmental cost performance. Additionally, we conduct an experimental study on a building water supply system to demonstrate the empirical performance of LAOC. The results indicate that LAOC can effectively reduce environmental and energy costs while guaranteeing safety constraints.

CCS Concepts

• **Computing methodologies** → **Computational control theory**.

ACM Reference Format:

Jianyi Yang, Pengfei Li, Tongxin Li, Adam Wierman, and Shaolei Ren. 2025. Learning-Augmented Online Control for Decarbonizing Water Infrastructures. 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, 18 pages. <https://doi.org/10.1145/3679240.3734595>

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only. Request permissions from owner/author(s).

E-ENERGY '25, Rotterdam, Netherlands

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1125-1/25/06
<https://doi.org/10.1145/3679240.3734595>

1 Introduction

Water supply is a critical utility for numerous infrastructures, including residential and commercial buildings, manufacturing facilities, and data centers. Globally, water systems consume about 4% of the total electricity use [3]. In municipalities, energy consumption of water systems typically accounts for approximately 30% to 40% of the total electricity use [2]. In the United States alone, the energy costs associated with water infrastructure amount to around 4 billion annually and contribute over 45 million tons of greenhouse gases [2]. Pumping is usually the most energy-intensive part of water infrastructures, representing up to 80% of the energy consumed by municipal water systems [6]. This significant energy consumption has spurred widespread interest in optimizing water pump systems to reduce both greenhouse gas emissions and monetary costs [2, 3, 6].

In most critical infrastructures, water supply systems use storage tanks to ensure a reliable water provision. Pumps are employed to maintain adequate water levels of these tanks to meet water demand. Beyond providing a reliable water supply, these tanks can serve as buffers that can be exploited to manage pumping systems more efficiently, thereby reducing greenhouse gas emissions and monetary costs. With the integration of renewable energy, both carbon intensity and electricity prices fluctuate over time [52, 87]. This time-varying property, combined with the widespread deployment of sensors, allows water supply systems to dynamically schedule the activation and/or the speed of pumps with the goal of optimizing carbon/energy efficiency [19, 75]. Importantly, the scheduling policy should ensure safe water levels of the tanks to address any emergencies.

Water supply management is an online control problem characterized by time-varying dynamics and cost functions that are revealed sequentially to the pump controller. Such problems are challenging due to the uncertainty of future contexts including demand, carbon intensity, and/or energy prices [9, 27, 56, 97]. Without precise knowledge of the future contexts, the controllers of pumping systems are difficult to achieve high energy efficiency. Nevertheless, exploiting the data of water usage, carbon intensity and energy price, machine learning (ML) can be applied to overcome the uncertainties inherent in online control, often surpassing

the performance of manually designed policies [54, 58, 60, 61]. Recently, ML predictions have been utilized in water supply systems to enhance cost savings and carbon efficiency [19, 93, 94].

However, ML can sometimes provide inaccurate predictions or low-quality advice, which can lead to arbitrarily poor performance and raise safety concerns for critical water infrastructures. For instance, a water tank in a conference center is crucial for ensuring a reliable water supply and fire protection. If the controller fails to maintain a safe water level, serious accidents can occur in the event of a municipal distribution system fault or a fire emergency. Naive deployments of ML-based controllers could result in such failures, leading to significant safety risks. Despite significant efforts to improve ML models for water supply systems [19, 83, 84], ML-based controllers fundamentally lack performance guarantees, especially for adversarial or out-of-distribution problem instances. Such lack of performance guarantees hinders the deployment of ML in real-world critical infrastructures.

To solve the fundamental challenges of ensuring worst-case performance guarantees for ML-based controllers, we propose a method that leverages control priors. Control priors are human-crafted online algorithms with provable worst-case performance guarantees [39, 43, 78, 79] or trusted rule-based heuristics that have been reliably used in real systems for a long time [21, 69]. These control priors are highly reliable in terms of safety metrics. By integrating these priors into ML-based controllers, we aim to develop an algorithm that ensures the safety performance of the ML-based controller is no worse than a the safety performance benchmark. Drawing on the concept of learning-augmented algorithms that incorporate ML advice into algorithm design, we call our proposed algorithm Learning-Augmented Online Control (LAOC).

While initially developed for water systems, the proposed algorithm (LAOC) is versatile and can be applied to various practical online control and resource management problems, such as battery management for electric vehicle (EV) charging station [86], workload scheduling for sustainable data centers [77], and control of cooling systems [69]. Adaptation of LAOC to these applications can improve the average performance while providing a worst-case performance guarantee.

Contributions. The contributions of the paper are summarized as follows. First, it presents an online control framework designed to sustainably and safely manage water supply for critical infrastructures. The framework addresses the urgent need for a worst-case safety risk guarantee in decarbonizing critical infrastructures. Notably, this framework extends to various online control and resource management problems across different critical infrastructures. Central to the paper’s contribution is the development of a novel learning-augmented algorithm named LAOC, which integrates a control prior into the ML-based controller to ensure worst-case safety risk constraints while optimizing decarbonization performance. Our analysis demonstrates that the proposed method reliably satisfies safety performance constraints for any problem instance while effectively leveraging ML predictions for decarbonization and cost saving. Furthermore, our analysis illuminates the tradeoff between the decarbonization and cost saving performance and the worst-case safety guarantee. Lastly, the paper evaluates the proposed algorithm for the water supply system

of critical buildings. Results indicate that LAOC achieves significant carbon reduction and cost savings compared to traditional controllers used in water supply systems focusing on maintaining water levels. Moreover, it showcases the advantage of LAOC in guaranteeing worst-case safety performance compared to pure ML-based algorithms.

2 Related Work

Optimization of water supply systems. The considered problem stems from the tradition field of water supply management. In this area, a lot of works consider the scheduling for water distribution systems [28, 71, 81, 85]. Some works have developed the pump control methods to maintain a water level for demand satisfaction and save energy, which has been studied in [19, 34, 68, 75, 83, 84, 93, 94]. Most of these works only consider the energy price, but do not explicitly consider the dynamical carbon intensity. The carbon emission of water infrastructures has recently become a crucial social concern [1, 2], so we include the carbon emissions in the optimization objective to ensure sustainable operation.

Much of the literature, e.g., [19, 93, 94], utilizes ML predictions of the future demand and/or energy price to improve the control performance. To fight against the future uncertainty, some works have developed robust control algorithms or constrained control algorithms for water supply systems [35, 44, 81, 85]. They either satisfy the safety constraints with a large probability or provide no guarantee on safety constraints. However, it is critically needed for water infrastructures to guarantee the worst-case safety performance of water supply given any problem instance. In this paper, we solve this challenge by designing a novel learning-augmented control algorithm utilizing the trusted control prior.

Online control. Our problem formulation is relevant to the literature of online competitive control. In our problem setting, the target is to minimize the cumulative cost in the nonlinear dynamics, which is different from the traditional control literature that uses measures for stabilization purposes [31, 32, 51, 74]. Like the recent works on competitive control [38, 40, 41, 43, 72, 79, 101], our work considers guarantees on the worst-case competitiveness, but our main focus is different — we leverage ML to explore policies with low average cost while enforcing competitiveness guarantees for any step in any episode. This enables the use of the existing competitive control policies as priors. Achieving our objective requires novel design of safe action sets and new analysis techniques to find the trade-off between the average performance and worst-case competitiveness.

Learning-based online control. Our algorithm is relevant to the broad area of learning-based control [16, 30, 33, 46, 59, 62, 76, 88]. These works have developed machine learning models to predict the system dynamic or control-relevant information which is utilized in deciding the control actions [13, 30, 33, 48, 49, 62, 88]. Recent works combine learning-based methods with system models in order to improve the safety or robustness of learning for control [16, 30, 59, 76, 89, 92]. Among them, learning-augmented online algorithms combine potentially untrusted ML predictions with robust policies (i.e., control priors). Learning-augmented algorithms have been developed for online control/optimization by combining ML predictions and control priors through online switching [12, 76]

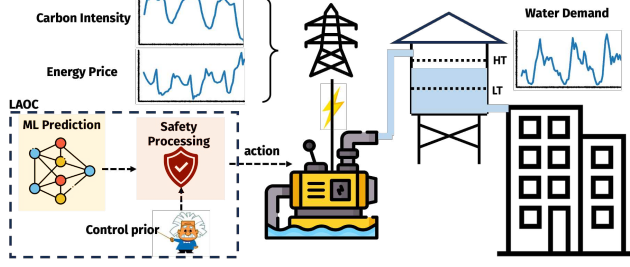


Figure 1: Water Supply Infrastructure with ML predictions.

or adaptively setting a confidence on the ML prediction [58, 59]. Compared to these studies, we make contributions by considering a more challenging setting, i.e., non-linear and time-varying dynamic models that are sequentially revealed online. Although some of the existing studies [23, 55, 57, 59, 76] provide provable cost bounds, they cannot guarantee a flexible any-step safety constraint given an arbitrary control prior, but this is needed for real problems [69].

Safe/Constrained Reinforcement Learning Our algorithm is also relevant to the literature of safe/constrained Reinforcement Learning (RL). Some safe/constrained RL works focus on discrete actions and their regret scales with the size of action set [29, 66, 95] while others [7, 98, 99, 99] apply to the continuous control problems. However, most of them only satisfy the constraints in expectation or with a high probability [7, 10, 18, 25, 26, 29, 36, 98, 99, 99]. A recent work [82] tries to solve RL with safety constraints satisfied almost surely, but no theoretical constraint satisfaction is guaranteed. When these algorithms are applied to online control systems like water supply management, the safety constraints can still be violated for some adversarial sequences. By contrast, our algorithm exploits the control priors and provides a theoretical guarantee for the safety constraint satisfaction.

3 Problem Formulation

In this section, with the water supply management as the application scenario, we present the safe online control model. Next, we show the safe online control model applies to broader applications by specifying the dynamics, loss and risk functions. Finally, we give the assumptions on the dynamics and risk functions required for the algorithm design and analysis.

3.1 Safe Online Control for Water Supply

In this section, we formulate an online control problem with time-varying costs and dynamics that captures the task of water supply management. A problem instance consists of H time slots. At the beginning of each time slot $h \in [H]$, the controller observes the water level state x_h and decides on an action $u_h \in \mathbb{R}^d$ to schedule the activation time and/or the speed of the pumps. This action incurs a non-negative carbon emission $c_e(u_h, e_h)$ related to the carbon intensity e_h , and a monetary cost $c_p(u_h, p_h)$ related to the energy price p_h . Given the water level state x_h and the action u_h , the system transitions to x_{h+1} at the end of slot h following the dynamic function f defined as:

$$x_{h+1} = f_h(x_h, u_h) = x_h + g(u_h) - w_h, \quad h = 1, \dots, H, \quad (1)$$

where w_h is the water consumption within time slot h , and g maps the control signal u_h to the amount of water supply within time slot h . Note that $g(u_h)$ is a linear function if we only control the activation time of pumping, and it is a nonlinear continuous function if we control the speed of pumps [90]. The water level state is expected to remain close to a nominal water level \bar{x} in the water tanks. Deviation from this nominal water level incurs a penalty cost denoted as $c_w(x_h)$.

For convenience, we denote $y_h := (e_h, p_h, w_h)$, and so $y_{1:H} = (y_1, \dots, y_H)$ is the information for the entire episode. The total loss at slot h is expressed as:

$$c_h(x_h, u_h) = \gamma_1 \cdot c_w(x_h) + \gamma_2 \cdot c_e(u_h, e_h) + \gamma_3 \cdot c_p(u_h, p_h), \quad (2)$$

where γ_1 , γ_2 , and γ_3 are weights used to convert the costs to the same measurement. An online control policy, denoted by π , outputs the action u_h . The cumulative loss within an episode of H time slots, following policy π , is expressed as: The offline optimal loss is denoted by J_H^* .

Safety Constraint. Online control algorithms must guarantee safety performance. For critical infrastructures, water supply management should maintain a safe water level to ensure reliable supply during emergencies. Failure to maintain a safe water level incurs a safety risk, denoted as $r_h(x_h, u_h)$. Given a nominal water level \bar{x} , a concrete form of safety risk can be denoted as

$$r_h(x_h, u_h) = \gamma_w \cdot \text{dist}(x_h, \bar{x}) + \gamma_b \cdot b(u_h), \quad (3)$$

where $\text{dist}(x_h, \bar{x})$ is a measure of distance between the water level x_h and the nominal water level \bar{x} , $b(u_h)$ penalizes the power load of the scheduling action u_h , and γ_w and γ_b are balancing weights for the two risk metrics. Note that the distance function $\text{dist}(x_h, \bar{x})$ can be an asymmetric function which provides different penalties for $x_h - \bar{x} \leq 0$ and $x_h - \bar{x} > 0$. The asymmetric distance measure is flexible to model different penalties of overly-high and overly-low water levels. We define the total safety risk of a policy π over an episode with H rounds as $R_H^\pi = \sum_{h=1}^H r_h(x_h, u_h)$.

To evaluate whether a controller is safe, we require a safety benchmark. In this paper, we use the scaled safety risk of an existing safe control prior π^\dagger as our benchmark. This means that for any problem instance $y_{1:h}$ and $h \in [H]$, the controller π must satisfy the safety constraint expressed as:

$$R_h^\pi \leq (1 + \lambda) R_h^{\pi^\dagger}, \quad (4)$$

where $R_h^{\pi^\dagger}$ is the safety risk of the safe control prior π^\dagger and $\lambda > 0$ is a preset parameter indicating the safety requirement level. The constraint in (4) is called $(1 + \lambda)$ -safety.

The intuition behind the safety constraint is that if the control prior has a worst-case safety performance guarantee for any instance, then a policy satisfying this constraint also ensures a performance guarantee adjustable by $1 + \lambda$. This constraint must be satisfied in each round to provide a strong worst-case guarantee. The safe control prior can be a human-crafted algorithm with a theoretical worst-case performance guarantee or a reliable heuristic implemented in real systems for a long time. In water infrastructures, the control prior can be a traditional controller that is designed to maintain the safe water level [93, 94].

Objective. We exploit ML predictions to optimize the expected loss while guaranteeing safety constraint for any problem instance.

Given a safety requirement $\lambda > 0$, the objective is:

$$\begin{aligned} & \min_{\pi} \mathbb{E}_{y_{1:H}} [J_H^{\pi}] \\ & \text{s.t. } R_h^{\pi} \leq (1 + \lambda)R_h^{\pi^{\dagger}}, \quad \forall h \in [H], \quad \forall y_{1:H} \in \mathcal{Y}. \end{aligned} \quad (5)$$

For convenience, we define the collection of all control policies that satisfies the safety requirement with λ as $\Pi_{\lambda} = \{\pi \mid R_h^{\pi} \leq (1 + \lambda)R_h^{\pi^{\dagger}}, \forall h \in [H], \forall y_{1:H} \in \mathcal{Y}\}$. If λ is larger, then the size of Π_{λ} is also larger, providing more flexibility to optimize the average loss. To solve this objective, we need to integrate the control prior π^{\dagger} into the ML-based controller.

3.2 Broader Applications

While the formulation is specifically given for water supply management, it applies to many other online control problems by replacing the dynamic function f_h , the cost function c_h and the risk function r_h with concrete expressions. Here, we give the following two application examples.

- **Battery management of EV charging station.** The battery management of Electrical Vehicle (EV) charging station is an online control problem where the agent needs to decide the amount of battery charging or discharging u_h at each round to maintain a nominal State of Charge (SoC) x_h that satisfies the charging demand [53, 59]. In this problem, the dynamic of SoC x_h is modeled by the dynamic function $f_h(x_h, u_h) = x_h + u_h - w_h$ where w_h is the charging demand, the loss function c_h defines the cost of charging and discharging. The risk function r_h defines the risk of not satisfying the charging demand. Classic controllers [11, 100] can serve as the control prior π^{\dagger} with risk performance guarantee.
- **Cooling control for sustainable data centers.** In this application, the target of the data center agent is to maintain a temperature range with high carbon efficiency by making online decisions of cooling equipment management [22, 69, 96]. Failure to maintain a suitable temperature range will overheat the devices and render the risk of critical services denial. The dynamic function $f_h(x_h, u_h)$ models the temperature dynamic where w_h is the randomness factor affecting the temperature change. The cost function c_h captures the losses of carbon emission and energy costs. The risk function r_h measures the risk of deviating from the normal temperature range. The traditional rule-based heuristics [69] that have verified performance in maintaining a suitable temperature can serve as the control prior π^{\dagger} .

3.3 Assumptions

In this paper, we assume the following conditions on the dynamic functions and the risk functions.

Assumption 3.1 (Lipschitz dynamics). *For each time h , the function f_h is Lipschitz continuous with respect to x_h and u_h with Lipschitz constants $\sigma_x \geq 0$ and $\sigma_u \geq 0$, respectively, i.e., for any (x, u) and (x', u') , f_h satisfies*

$$\begin{aligned} \|f_h(x, u) - f_h(x', u)\| & \leq \sigma_x \|x - x'\| \\ \|f_h(x, u) - f_h(x, u')\| & \leq \sigma_u \|u - u'\|. \end{aligned}$$

Assumption 3.2 (Well-conditioned risk functions). *For each time h , the risk function r_h is non-negative, α -strongly convex, and β -smooth with respect to (x_h, u_h) .*

The first assumption is the Lipschitz continuity of the dynamic functions, which is common in finite-horizon control models [59, 60, 101]. For water supply management, the dynamic function f_h in (1) is clearly Lipschitz continuous as g is a Lipschitz continuous function.

The second assumption is the non-negativity, convexity and smoothness of the risk functions, which is a common regularity condition in control system costs [60, 64, 65, 78]. We are flexible to choose different risk functions that satisfy Assumption 3.2. For example, we can choose an asymmetric dist function as $\text{dist}(x_h, \bar{x}) = \gamma_{w,1}(\bar{x} - x_h)^2$ if $\bar{x} \geq x_h$ and $\text{dist}(x_h, \bar{x}) = \gamma_{w,2}(x_h - \bar{x})^2$ if $\bar{x} < x_h$ and a quadratic penalty of the power load, and the obtained risk function satisfies Assumption 3.2.

4 Learning-Augmented Online Control (LAOC)

In this section, we present and analyze an algorithm, LAOC, to solve the online control problem introduced in the previous section. Before stating the algorithm, we highlight the challenges created by the safety requirements.

4.1 Challenges Due to Safety Requirements

Our goal is to find a policy satisfying safety constraints in (5) while exploiting the ML predictions to achieve a low loss. However, this is very challenging for online control where the future contexts are unknown to the agent.

One might hope that a straightforward design that considers a linear combination of the control prior and the pure ML action (Lin) would be sufficient. Formally, Lin is defined as $\pi = \rho\tilde{\pi} + (1 - \rho)\pi^{\dagger}$, where $\tilde{\pi}$ is the pure ML policy and π^{\dagger} is the policy prior. However, Proposition 4.1 shows that, unless we completely ignore the ML policy (i.e., $\rho = 0$), Lin cannot guarantee $(1 + \lambda)$ -safety given any ML policy.

Proposition 4.1. *Define the quality of pure ML as the normalized difference between the ML advice and the offline optimal action $\|\tilde{u} - u^*\|^2 / J_H^*$. If the pure ML have an arbitrarily low quality (i.e., $\|\tilde{u} - u^*\|^2 / J_H^* \rightarrow \infty$), Lin with $\rho \in (0, 1]$ cannot guarantee $(1 + \lambda)$ -safety for any finite $\lambda > 0$.*

Proposition 4.1 is proven by constructing a contradictory example that if $(1 + \lambda)$ -safety with finite λ is satisfied by Lin with $\rho \in (0, 1]$, the quality of ML advice $\frac{\|\tilde{u} - u^*\|^2}{J_H^*}$ must be bounded by a finite value. In other words, $(1 + \lambda)$ -safety cannot be satisfied by Lin with a potentially unsafe ML model in the worst case.

Overcoming the limitation of Lin with respect to safety guarantees requires a more flexible combination of pure ML and the control prior. Thus, we give a second natural approach maps the ML advice into a safe action set defined by the safety constraint for each round h as

$$\overline{u}_{\lambda,h} = \left\{ u_h \mid R_h \leq (1 + \lambda)R_h^{\pi^{\dagger}} \right\}, \quad (6)$$

where $R_h = \sum_{i=0}^h r_i(x_i, u_i)$ and $R_h^{\pi^{\dagger}} = \sum_{i=0}^h r_i(x_i^{\dagger}, u_i^{\dagger})$. The mapping can be a linear combination that selects the action as $u_h = \bar{\rho}_h \tilde{u}_h +$

$(1 - \bar{\rho}_h)u_h^\dagger$ where $\bar{\rho}_h = 1$ if $\tilde{u}_h \in \overline{\mathcal{U}}_{\lambda,h}$ and $\bar{\rho}_h$ is the solution of $r_h(x_h, \bar{\rho}_h \tilde{u}_h + (1 - \bar{\rho}_h)u_h^\dagger) = (1 + \lambda)R_h^{\pi^\dagger} - R_{h-1}$ if $\tilde{u}_h \notin \overline{\mathcal{U}}_{\lambda,h}$. We refer to this policy as `Lin+`.

`Lin+` uses a time-varying combination variable $\bar{\rho}_h$, so it is much more flexible than `Lin` and can strictly guarantee $(1 + \lambda)$ -safety given any instance as long as the safe action set in (6) is non-empty. Unfortunately, the naive design of safe action set $\overline{\mathcal{U}}_{\lambda,h}$ in (6) can be empty, which results in no feasible actions. This is illustrated by the following example.

Example 4.1. Suppose that $\sum_{i=0}^h r_i(x_i, u_i) = (1 + \lambda) \sum_{i=0}^h r_i(x_i^\dagger, u_i^\dagger)$ is satisfied at time h . If $x_{h+1} = x_{h+1}^\dagger$ holds at round $h + 1$, the agent can always choose $u_{h+1} = u_{h+1}^\dagger$ to satisfy (6) at round $h + 1$. However, when $x_{h+1} \neq x_{h+1}^\dagger$, it is possible that the control prior has a low loss for its state x_{h+1} at time $h + 1$ such that for any action $u \in \mathcal{U}$ the true loss $c_{h+1}(x_{h+1}, u)$ is larger than the scaled prior loss $(1 + \lambda)c_{h+1}(x_{h+1}^\dagger, u_{h+1}^\dagger)$. In such a case, the naive safe action set $\overline{\mathcal{U}}_{\lambda,h+1}$ is empty, and the control agent cannot maintain the inequality in (6), thus potentially violating the subsequent safety constraints.

The failures of the intuitive policies `Lin` and `Lin+` show that for a policy to combine the ML advice and the control prior, it must be flexible and conservative enough to guarantee that feasible actions exist to meet the safety constraints. In the next section, we give the design that can theoretically guarantee the $(1 + \lambda)$ -safety for any sequence and ML advice.

4.2 Algorithm Design

In this section, we give an overview of the design of Learning Augmented Online Control (LAOC).

First, we highlight the design of the safe action set used in the algorithm. Instead of directly guaranteeing the inequality in (6), we ensure that the resulting cumulative loss satisfies $\sum_{i=0}^h r_i(x_i, u_i) + \phi_h \leq (1 + \lambda) \sum_{i=0}^h r_i(x_i^\dagger, u_i^\dagger)$ with an added reservation $\phi_h \geq 0$ for hedging. With a proper design of the reservation, $\mathcal{U}_{\lambda,h}$, $h \in [h, H]$ can be guaranteed to be not empty for all the possible future control environments $y_{h:H}$. To this end, we design a reservation in the next proposition, whose proof is deferred to Appendix C.

Proposition 4.2. Define a safe set $\mathcal{U}_{\lambda,h}$, $\lambda > 0$ as

$$\mathcal{U}_{\lambda,h} := \left\{ u_h \in \mathcal{U} \mid R_h + \phi_h(u_h) \leq (1 + \lambda)R_h^{\pi^\dagger} \right\}, \quad (7)$$

where $R_h = \sum_{i=0}^h r_i(x_i, u_i)$ and $R_h^{\pi^\dagger} = \sum_{i=0}^h r_i(x_i^\dagger, u_i^\dagger)$ are the true loss and the loss of control prior, respectively. Moreover, $\phi_h(u) = q_h \|x_{h+1} - x_{h+1}^\dagger\|^2 = q_h \|f_h(x_h, u) - f_h(x_h^\dagger, u_h^\dagger)\|^2$ is a reservation function, where $q_h = C_1(1 + \frac{1}{\lambda}) \frac{\beta}{2} \sum_{h'=0}^{H-h-1} (C_2 \sigma_x^2)^{h'}$ for constants $C_1 \geq 1$ and $C_2 \geq 1$. With Assumptions 3.1 and 3.2, if $\mathcal{U}_{\lambda,h-1}$ is not empty and the action u_{h-1} at round $h - 1$ is selected from the safe set $\mathcal{U}_{\lambda,h-1}$, then $\mathcal{U}_{\lambda,h}$ is not empty and always includes u_h^\dagger .

The key insight behind the formulation of the reservation $\phi_h(u)$ in (7) is to hedge against the possible violation of safety constraints in future rounds. If the resulting state difference $\|x_{h+1} - x_{h+1}^\dagger\|^2$ from choosing u_h is greater, the possible loss difference $\sum_{i=h+1}^H r_i(x_i, u_i) - (1 + \lambda)r_i(x_i^\dagger, u_i^\dagger)$ in the following rounds can also be greater in the worst case. Thus, the reservation $\phi_h(u)$ is designed as the scaled

Algorithm 1 Learning Augmented Online Control (LAOC)

Input: ML model $\tilde{\pi}$ and control prior π^\dagger

- 1: **for** time horizon $h = 0, \dots, H - 1$ **do**
 - 2: Observe state x_h , information $\{r_h, f_h\}$, and last-step context w_{h-1} .
 - 3: Update the policy prior's state $x_h^\dagger = f_{h-1}(x_{h-1}^\dagger, u_{h-1}^\dagger) + w_{h-1}$
 - 4: Obtain an action u_h^\dagger by the prior π^\dagger , and update prior risk $R_h^{\pi^\dagger} = R_{h-1}^{\pi^\dagger} + r_h(x_h^\dagger, u_h^\dagger)$
 - 5: Obtain the ML action \tilde{u}_h via the ML model $\tilde{\pi}$
 - 6: **if** the ML action $\tilde{u}_h \in \mathcal{U}_{\lambda,h}$ **then** take $u_h = \tilde{u}_h$
 - 7: **else** take $u_h = m(\tilde{u}_h)$ **end if** // Map to a safe action set $\mathcal{U}_{\lambda,h}$ (7) by (8) or (9)
 - 8: Update true loss $J_h = J_{h-1} + c_h(x_h, u_h)$ and risk $R_h = R_{h-1} + r_h(x_h, u_h)$
 - 9: **end for**
-

state difference to account for the worst-case future risk difference between the true control policy and the control prior π^\dagger .

As a consequence of Proposition 4.2, if u_h is selected from $\mathcal{U}_{\lambda,h}$ for each round h , there always exists a non-empty safe action set $\mathcal{U}_{\lambda,h}$ in the subsequent steps, and thus $(1 + \lambda)$ -safety is strictly satisfied for each round. Based on Proposition 4.2, given an ML policy $\tilde{\pi}$ and a control prior π^\dagger , we design the online learning-augmented control policy LAOC as shown in Algorithm 1. At each round h within an episode, the controller first evaluates the loss of the control prior. To achieve this, after observing the true state x_h and f_{h-1}, w_{h-1} , we first calculate a "virtual state" corresponding to the control prior for the same online information $y_{0:h-1}$, denoted by $x_h^\dagger = f_{h-1}(x_{h-1}^\dagger, u_{h-1}^\dagger) + w_{h-1}$. Next, we query the control prior π^\dagger with a state x_h^\dagger and obtain an action u_h^\dagger , which can be used to update the cumulative risk $R_h^{\pi^\dagger}$ at round h . By doing so, a safe action set $\mathcal{U}_{\lambda,h}$ can be constructed by Proposition 4.2.

To utilize the ML advice for loss performance, we select an action that is close enough to the pure ML action from the safe action set. If the ML action \tilde{u}_h is in the safe action set $\mathcal{U}_{\lambda,h}$, then we simply select $u_h = \tilde{u}_h$. Otherwise, we can use a mapping function $m: \mathbb{R}^d \rightarrow \mathcal{U}_{\lambda,h}$ that maps the ML action \tilde{u}_h into an action in the safe action set. One choice of m is the projection operation which selects action as

$$u_h = m(\tilde{u}_h, \mathcal{U}_{\lambda,h}) = \arg \min_{u \in \mathcal{U}_{\lambda,h}} \|\tilde{u}_h - u\|. \quad (8)$$

When the safe action set is a convex set (e.g. the dynamic functions $\{f_h : h \in [H]\}$ are linear [38, 40, 101, 103]), the projection can be efficiently solved. Otherwise, the complexity can be high especially for high dimensional actions [20, 63]. Under such cases, we can choose m as a linear combination as below

$$u_h = m(\tilde{u}_h, \mathcal{U}_{\lambda,h}) = \rho_h \tilde{u}_h + (1 - \rho_h)u_h^\dagger, \quad (9)$$

where we need to solve an one-dimensional combination variable $\rho_h \in [0, 1]$ as a solution of $R_{h-1} + r_h(x_h, \rho_h \tilde{u}_h + (1 - \rho_h)u_h^\dagger) + \phi_h(\rho_h \tilde{u}_h + (1 - \rho_h)u_h^\dagger) = (1 + \lambda)R_h^{\pi^\dagger}$. We will prove in Theorem 4.4 that LAOC with both mapping functions in (8) and (9) share the same expected loss bound.

The time complexity of LAOC is $O(H(T_{\text{ML}} + T_{\text{prior}} + T_{\text{map}}))$ where T_{ML} , T_{prior} and T_{map} are the time complexities of the ML inference, the control prior and the mapping operations, respectively. T_{ML} is determined by the ML architecture. The time complexity of the control prior T_{prior} usually increases with the complexity of the control problem. Take the control prior ROBD [41] as an example, the complexity to solve the optimization in ROBD scales with the dimension of the action. Furthermore, the mapping complexity T_{map} depends on the action-state dimensions and the complexity of the control model. If the safe action set in (7) is convex (e.g. linear dynamic leads to a convex safe action set), we can use a convex optimization solver to efficiently solve the projection in (8). When the safe action set is non-convex, the projection into a non-convex action set has a high time complexity. In such cases, we can map the ML action to the safe action set by solving an one-dimensional combination variable in (9).

Safety-aware finetuning. If we have access to the pure ML model, we can finetune it based on available sequence data to further improve average loss performance with the safety guarantee. Specifically, given the pure ML model $\tilde{\pi}$ which outputs the ML action \tilde{u}_h and the safe action set $\mathcal{U}_{\lambda,h}$, we finetune the ML model by minimizing the empirical loss of safe actions:

$$\tilde{\pi}_\lambda^{(n)} = \arg \min_{\tilde{\pi} \in \Pi} \sum_{y_{1:H} \in \mathcal{D}_n} \sum_{h=1}^H c_h(x_h, m(\tilde{u}_h, \mathcal{U}_{\lambda,h})), \quad (10)$$

where \mathcal{D}_n is the finetuning dataset with n sequences. To finetune the ML model with (10), we can directly perform the back-propagation through the online process where all the operations are differentiable. The projection in (8) can be implicitly differentiated as shown in [8]. The linear mapping in (9) is also differentiable by differentiating the equation to solve ρ_h .

4.3 Performance Bounds

We provide the performance analysis of LAOC in this section. We first present the conclusion that the the safety constraint in (4) is always satisfied by LAOC. Next, we give the average performance bound of LAOC under the safety guarantee. Last but not least, we provide the performance bound by safety-aware finetuning in Eqn.(10).

4.3.1 Safety constraint satisfaction. In Proposition 4.2, we prove that the safe set $\mathcal{U}_{\lambda,h}$ in (7) is not empty for each round h . Since LAOC (Algorithm 1) guarantees that the action u_h lies in the safe set at each round, we can get the conclusion of safety constraint satisfaction in the next theorem.

Theorem 4.3. *By LAOC (Algorithm 1) with safety set $\mathcal{U}_{\lambda,h}$ in (7), for any problem sequence $y_{1:H}$ and any round $h \in [H]$, we can guarantee that the safety risk constraint in (4) is satisfied.*

Theorem (4.3) highlights that LAOC can strictly guarantee $(1 + \lambda)$ -safety for any problem instance even when the ML policy $\tilde{\pi}$ has an arbitrarily bad performance. Under the safety guarantee, we are concerned about the expected loss performance given in the next theorem.

4.3.2 Average performance. The expected loss relies heavily on the choices of C_1 and C_2 in (7) of Proposition 4.2. To see this, if C_1 or C_2 is larger, the reservation $\phi_h(u)$ becomes larger, so the safe action set

$\mathcal{U}_{\lambda,h}$ contains less feasible actions. Thus, the policy cannot utilize the ML model to improve the average loss performance effectively. On the contrary, if C_1 and C_2 approach 1, it can happen in the earlier rounds that the sizes of the safe action sets $\mathcal{U}_{\lambda,h}$ are too large and the selected action is too far from the prior action. This results in large state differences between x_h and x_h^\dagger in future rounds, resulting in small safe action set $\mathcal{U}_{\lambda,h}$ and impeding the exploitation of ML advice. The following analysis formally shows the factors that affect the expected performance and suggest the choices of C_1 and C_2 .

Theorem 4.4. *Assume that the ML policy $\tilde{\pi}$ is L_π -Lipschitz continuous and the function c_h is L_c -Lipschitz continuous, by optimally choosing $C_1 = 1 + \frac{1}{\sqrt{1+\lambda}}$ and $C_2 = \arg \min_{c \geq 1} \{ \frac{c}{c-1} \sigma_u^2 (1 - (c\sigma_x^2)^{H-h}) / (1 - c\sigma_x^2) \}$ in (7), the expected loss of LAOC π_λ that guarantees $(1 + \lambda)$ -safety is bounded by*

$$\mathbb{E} [J_H^{\pi_\lambda}] \leq \mathbb{E} [J_H^{\tilde{\pi}}] + B \mathbb{E} \left[\sum_{h=0}^{H-1} \left[\delta_h - (\sqrt{1+\lambda} - 1)^2 Gr_h^\dagger \right]^+ \right], \quad (11)$$

where $\delta_h = \|\tilde{\pi}(\tilde{s}_h) - u_h^\dagger\|$ is the action discrepancy between the pure ML action and the control prior, $G := 2(L_c(1 + \frac{C_2}{C_2-1}\sigma_u^2(1 - (C_2\sigma_x^2)^{H-h})/(1 - C_2\sigma_x^2)))^{-1}$ and $B := L_c(1 + (1 + 2L_\pi)\sigma_u \sum_{i=0}^{H-1} (\sigma_x + 2\sigma_u L_\pi)^{h-i-1})$ are constants of the control system, in which β is the smoothness parameter of the risk function r_h , A is the size of the state-action set, L_π is the Lipschitz constant of the ML advice policy $\tilde{\pi}$, σ_x and σ_u are the Lipschitz constants of the dynamics model f_h .

The expected loss bound in Theorem 4.4 relies on the choices of C_1 and C_2 . When λ becomes larger, the safety constraint is more relaxed, so a smaller C_1 is chosen to get a smaller reservation $\phi_u(h)$ in Proposition 4.2, allowing more flexibility to follow the ML advice. Also, C_2 is selected to alleviate the impact of the dynamic sensitivity measured by σ_x and σ_u (Assumption 3.1) on the expected loss.

The expected loss bound in Theorem 4.4 can be interpreted as follows. First, the safety constraint naturally creates a gap of expected loss between LAOC π_λ and the ML advice $\tilde{\pi}$. More specifically, given a control prior π^\dagger , when $\lambda > 0$ becomes smaller, the safety constraint is more stringent, which thus makes the actions of LAOC π_λ potentially deviate more from those of the ML advice policy $\tilde{\pi}$ and increases the bound in (11). On the contrary, when $\lambda > 0$ becomes larger, the safety constraint is more relaxed, reducing the expected loss of LAOC π_λ . In particular, if λ is sufficiently large, the term $[\delta_h - (\sqrt{1+\lambda} - 1)^2 Gr_h^\dagger]^+$ can reduce to zero, voiding the safety constraint and resulting in the same expected loss as pure ML. Additionally, the expected loss is affected by the action discrepancy δ_h because a larger δ_h means larger difference between the prior π^\dagger and the ML model $\tilde{\pi}$, naturally making it more difficult for LAOC to approach ML $\tilde{\pi}$ while satisfying safety constraints.

4.3.3 Generalization performance of safety-aware finetuning. In this section, we consider the case in which the ML policy in LAOC is trained on (10). We bound the average loss gap between LAOC policy and the unconstrained-optimal policy π^* . We denote $\pi_\lambda^{(n)}(s_h) = m(\tilde{\pi}_\lambda^{(n)}(s_h), \mathcal{U}_{\lambda,t})$ as LAOC policy by Algorithm 1 with the ML model $\tilde{\pi}_\lambda^{(n)}$ trained by (10) on a dataset with n traces, and bound the expected loss $\mathbb{E}[J_H^{\pi_\lambda^{(n)}}]$ in the following theorem.

Theorem 4.5. *If ML policy is trained by the loss function in Eqn. (10) with a training dataset with n samples, with probability at least $1 - \delta$, $\delta \in (0, 1)$, the expected loss of our competitiveness-constrained policy $\pi_\lambda^{(n)}$ is bounded by*

$$\mathbb{E} \left[J_H^{\pi_\lambda^{(n)}} \right] \leq \mathbb{E} \left[J_H^{\pi^*} \right] + B \mathbb{E} \left[\sum_{h=0}^{H-1} \left[\delta_h - (\sqrt{1+\lambda} - 1)^2 G r_h^\dagger \right]^+ \right] + \mathcal{O} \left(\sqrt{\frac{1}{n} \ln \frac{N(\epsilon, \Pi_\lambda, \hat{L}_1^n)}{\delta}} \right),$$

where the system-related parameters B, G and δ_h have the same definition as in Theorem 4.4, $N(\epsilon, \Pi_\lambda, \hat{L}_1^n)$ is the ϵ -covering number of the competitive policy space Π_λ with L_1 -norm as the distance measure (the distance of two policies π and π' is $\|\pi - \pi'\|_{\hat{L}_1^n} = \frac{1}{n} \sum_{t=1}^n \sum_{h=1}^H \|\pi(s_h^{(t)}) - \pi'(s_h^{(t)})\|_1$) on the training dataset \mathcal{D}_n , and \mathcal{O} indicates the scaling with the loss upper bound P , the horizon H , and the size of action-state space $\mathcal{X} \times \mathcal{U}$.

Theorem 4.5 shows that as the number of training samples $n \rightarrow \infty$, the expected loss is bounded by the unconstrained-optimal expected loss $\mathbb{E} \left[J_H^{\pi^*} \right]$ plus an additional term relying on the expected loss of the prior π^\dagger and the parameter $\lambda > 0$. This additional term is because the policy is optimized under the safety constraint in (4). When λ becomes larger, the constraint is more relaxed and the expected loss is closer to the unconstrained-optimal expected loss. Also, Theorem 4.5 shows that our policy with the online-trained ML model converges with a rate of $\sqrt{1/n}$. In particular, the convergence rate is affected by λ through the covering number $N(\epsilon, \Pi_\lambda, \hat{L}_1^n)$ which indicates the richness of the competitive policy class Π_λ . Comparing to the unconstrained policy set Π_∞ , the covering number of the competitive policy class Π_λ is smaller. This is because with the same ML model, the safety constraint reduces the set of feasible actions — with a smaller $\lambda > 0$, the safe policy space becomes smaller, making it easier for the convergence of LAOC.

5 Case Study

In this section, we evaluate the performance of LAOC by experiments on a concrete water supply case and compare LAOC with different control baselines.

5.1 Setup

In this section, we provide the experimental setups on the water supply management. We first present the architecture of water supply system with roof top water tanks. Next, we introduce the datasets used in the experiments including the traces of water demand, carbon intensity, and energy price. Following that, we define the concerned performance metrics in the experiments. Finally, we provide the settings of LAOC and the baselines.

5.1.1 Water supply system with roof top water tanks. The water supply systems of many modern buildings are equipped with roof top water tanks. The roof top water tanks have large water storage capacities and exploit the gravity in the elevated level to supply water for building users. Water is pumped from municipal water sources to these roof top water tanks to maintain a water level. The

water tanks can play an important part in sustaining water supply system because the manager can pump less water (by decreasing the activation time and/or the speed of pumps) to the water tanks when the carbon intensity and energy price are high while still satisfying the demand using the water stored in the water tanks. Beyond that, the water tanks are crucial for the safety of the buildings because they are equipped to supply water for fire protection systems and the mission-critical functions of the buildings. Therefore, we must make sure that the water level in a water tank is not far from its nominal water level to meet the safety requirements.

To sustain the water supply system for a building with roof top water tanks, we need to know the energy consumption of its pumping system to pump water to the water tanks. In this paper, we estimate the power P_{pump} (kW) of water pumping by the following formula converted from the horsepower formula used in engineering practice [4]:

$$P_{\text{pump}} = \frac{\text{WF} \times \text{HD} \times \text{SG}}{102 \cdot \eta_{\text{pump}}}, \quad (12)$$

where WF(L/s) is the water volume flow, HD(m) is the height of the water tank, and SG = 1 is the water specific gravity, and η_{pump} is the power efficiency of the pumping system. We develop a controller that decides the amount of water u_h (m³) pumped into the water tanks in each hour round h . The effective water flow is $u_h/3.6$ (L/s)¹, which corresponds to an energy consumption of $(u_h \times \text{HD} \times \text{SG}) / (367.2 \cdot \eta_{\text{pump}})$ (in kWh) by (12). Here, we define the energy efficiency as the energy consumption to pump a unit m³ of water to the water tank in one hour and denote it as

$$\eta = \frac{\text{HD} \times \text{SG}}{367.2 \cdot \eta_{\text{pump}}}. \quad (13)$$

The setups in the experiment are given as below. The buildings are 75 m high and have water tanks with a total volume of 80m³ on the roof. Each control horizon has a span of 24 hours. By (13), the power consumption to pump a unit m³ of water is $\eta = 0.272$ kWh/m³ by choosing the energy efficiency of the pumps as $\eta_{\text{pump}} = 75\%$ according to [91]. The controller decides the amount of water pumped into the water tanks in each hour as u_h (m³), so the energy consumption at hour round h is $\eta \cdot u_h$ (kWh).

5.1.2 Performance Metrics. In water supply management, the concerned performance metrics include the carbon and energy costs and the safety risk. Given the water supply system with roof top water tanks, the expressions of the objectives are given as below.

Carbon cost. Given an action u_h which is the amount of pumped water at hour round h , the energy consumption is $\eta \cdot u_h$ (kWh). Therefore, given the carbon intensity e_h (g/kWh) at round h , the carbon emission at round h is $c_e(u_h, e_h) = e_h \cdot (\eta \cdot u_h)$.

Energy cost. With the energy price p_h for round h , the total energy cost at round h is $c_p(u_h, e_h) = p_h \cdot (\eta \cdot u_h)$.

Deviation from the safe level. We choose the nominal safe water level as $\bar{x} = 40\text{m}^3$ (half of the total water tank capacity). We choose a quadratic penalty for water level deviation which restrains large deviation. Thus, the deviation is measured by the quadratic deviation from the nominal level \bar{x} , i.e. $c_w(x_h) = (x_h - \bar{x})^2$.

¹The amount of water supply per hour can be adjusted by either controlling the activation time of the pumps or the speed of the pumps. For ease of computation, we assume the speed of the pumps is constant within each hour.

The loss function is a weighted combination of the deviation and the carbon and energy costs which is expressed as

$$c_h(x_h, u_h) = \gamma_1 \cdot (x_h - \bar{x})^2 + \gamma_2 \cdot e_h \cdot (\eta \cdot u_h) + \gamma_3 \cdot p_h \cdot (\eta \cdot u_h). \quad (14)$$

We consider the expected loss $\mathbb{E}_{y_{1:H}} [J_H^\pi] = \mathbb{E}_{y_{1:H}} \left[\sum_{h=1}^H c_h(x_h, u_h) \right]$ where the expectation is taken on the distribution of the water demand, carbon intensity and energy price traces.

Safety risk. The safety risk is determined by the deviation and the hourly energy consumption. A high deviation will increase the risk of not satisfying the water demand and a large energy consumption can add too much power load to the energy system. We consider a quadratic penalty to restrain large deviation and hourly energy consumption and the safety risk is expressed as

$$r_h(x_h, u_h) = \gamma_w \cdot (x_h - \bar{x})^2 + \gamma_b \cdot (\eta \cdot u_h)^2. \quad (15)$$

In some scenarios, we need to consider different penalties for overly-low and overly-high water levels and model the deviation as an asymmetric function. If the asymmetric function satisfies Assumption 3.2 (e.g. an asymmetric dist function as $\text{dist}(x_h, \bar{x}) = \gamma_{w,1}(\bar{x} - x_h)^2$ if $\bar{x} \geq x_h$ and $\text{dist}(x_h, \bar{x}) = \gamma_{w,2}(x_h - \bar{x})^2$ if $\bar{x} < x_h$), the theoretical conclusions of LAOC still hold, and the key observations of experiments also generalize to such asymmetric penalties.

Given a control prior π^\dagger , we consider $(1 + \lambda)$ -safety which guarantees for any sequence that the safety risk is always bounded by the scaled safety risk of π^\dagger , i.e. $\forall h \in [H], \forall y_{1:H} \in \mathcal{Y}, R_h^\pi \leq (1 + \lambda)R_h^{\pi^\dagger}$. We also directly evaluate the safety risk performance by the maximum risk ratio on the testing dataset $\max_{y_{1:H} \in \mathcal{D}_{\text{test}}} \left(R_H^\pi / R_H^{\pi^\dagger} \right)$, which is a commonly used metric for worst case performance. [40, 42, 78].

5.1.3 Water demand, carbon intensity and energy price. The experiments are conducted based on some public datasets. We provide the details for the traces of water demand, carbon intensity and energy price as below.

Water demand trace. The consumed water at each hour round h is w_h and affects the water level dynamics through (1). In our experiments, we use the water demand dataset measured for university buildings in [14]. For each building, the trace contains hourly water consumption from August 1st, 2018 to December 8th, 2018. Since the traces are measured on low-rise university buildings, we scale up the hourly water consumption by 10 to simulate the high-rise building with dense occupancy. The water consumption data of four residence hall is used for training the ML model for water supply management. We augment the water consumption data of another two residence halls and get the 1-year demand traces for 20 buildings which are held out for testing.

To further evaluate the robustness of the algorithms, we also create an Out-Of-Distribution (OOD) testing dataset on the basis of the original testing dataset. We generate the OOD demand dataset by adding Gaussian noise to each sample in the original dataset. The standard deviation of the Gaussian noise is set as 30% of the maximum demand value.

Carbon intensity trace The carbon intensity datasets are from California Independent System Operator (CAISO) which are published on the website of Electricity Maps [70]. The carbon intensity datasets contain the hourly carbon intensity of a city in California.

We use the carbon traces in 2022 to train the ML model, and we hold out the carbon traces in 2023 for validation and testing.

Energy price trace The electricity price datasets are from CAISO which are published on the website of Energy Online [5]. Each price trace in the dataset contains the energy price value every 5 mins. We convert the original traces into hourly price traces by calculating the average price within each hour. We use the price data in 2022 to train the ML model while holding out the price data in 2023 for validation and testing.

5.1.4 Settings of LAOC. To implement LAOC in Algorithm 1, we need an ML model $\tilde{\pi}$ and a control prior π^\dagger as inputs. Also, we can perform safety-aware finetuning in Eqn.(10) to learn an ML model for Algorithm 1. Thus, we summarize the variants of LAOC as follows.

- **LAOC ($\lambda, \tilde{\pi}, \pi^\dagger$):** We use ML model $\tilde{\pi}$ and control prior π^\dagger as the inputs of LAOC. $\tilde{\pi}$ and π^\dagger can be replaced with a concrete ML model and a concrete control prior, respectively. If not specified, LAOC uses Online Gradient Descent (OGD) as the control prior by default. If not specified, LAOC uses the ML model purely trained without considering the safety constraint by default. λ determines the $(1 + \lambda)$ -safety in (4).
- **LAOC-F(λ, π^\dagger):** We use control prior π^\dagger and an ML model obtained by safety-aware finetuning in Eqn.(10) as the inputs of LAOC. If not specified, LAOC-F uses Online Gradient Descent (OGD) as the control prior by default. λ determines the $(1 + \lambda)$ -safety in (4).

ML model. The ML model for LAOC is a recurrent neural network. It takes available information about demand, carbon intensity, and electricity price as inputs and outputs the action for each round. By default, the ML model has 2 hidden layers and each hidden layer has 12 neurons. The ML model is trained by the Adam optimizer with a learning rate 5×10^{-4} for 400 epochs.

Control prior. The control prior can be selected from some controllers that focus on reducing the safety risk. [93, 94]. Some robust online optimization algorithms such as Online Gradient Descent (OGD) [24], Online Balanced Descent (ROBD) [41] can be applied to optimize the safety risk, so they can serve as the control prior. Alternatively, we can apply Model Predictive Control (MPC) [93, 94] to minimize the safety risk which is commonly used for water supply control as a control prior.

Regarding the safe set in (7), the safety requirement parameter λ is chosen from $[0, 2]$. C_1 and C_2 are chosen based on Theorem 4.4.

5.1.5 Baselines. We compare LAOC with OGD [24], ROBD [41] and MPC [94] that focus on the safety risk, the pure ML that is trained on the average loss, and the naive learning-augmented design Lin.

- **Offline Optimal Policy (OPT):** This is the optimal offline policy that knows all the information in advance and obtains the optimal action for each episode.
- **Online Gradient Descent (OGD):** Online gradient descent [24] is an online algorithm to minimize the safety risk without relying on any predictions. OGD has provable regret bound and competitive ratio with proper choice of step size. We use OGD as a control policy prior by default.
- **Regularized Online Balanced Descent (ROBD):** ROBD is an online optimization algorithm to minimize the safety risk with one-step

Table 1: Cost and risk performance

Metrics	Control priors				ML		Learning-augmented designs			
	OGD	ROBD	MPC-LSTM	TMPC	ML	CRL	Lin-0.5	Lin-0.2	LAOC ($\lambda = 0.4$)	LAOC ($\lambda = 0.8$)
Avg. energy (US\$)	7344	7347	7008	7640	6169	6584	6169	7190	6872	6690
Avg. carbon (kg)	17994	18278	18007	18820	16123	16820	17178	17668	17037	16526
Max risk ratio	2.04	1.14	6.19	3.608	6.17	4.60	4.56	2.44	2.76	3.40

Table 2: Cost and risk performance under OOD setting

Metrics	Control priors				ML		Learning-augmented designs			
	OGD	ROBD	MPC-LSTM	TMPC	ML	CRL	Lin-0.5	Lin-0.2	LAOC ($\lambda = 0.4$)	LAOC ($\lambda = 0.8$)
Avg. energy (US\$)	7610	7580	7136	7937	6516	6854	7063	7391	6901	6798
Avg. carbon (kg)	20494	20577	18630	21143	18476	18711	19485	20090	19326	19060
Max risk ratio	4.09	1.24	41.2	12.1	11.00	7.68	5.79	3.32	4.36	5.68

demand prediction. It enjoys provable competitive ratio given perfect one-step prediction [41, 72]. We use ROBD as a control policy prior. We set the parameters for ROBD optimally according to [41].

- **Model Predictive Control (MPC):** MPC [17] solves the control problem by leveraging predictions of the future information. Here, we assume that at round h , the information $w_{h:H}$ is predicted as $\hat{w}_{h:H}$, and the per-round prediction error normalized by the maximum input range is $\epsilon = \mathbb{E} \left[\frac{1}{(H-h+1)(w_{\max})} \|w_{h:H} - \hat{w}_{h:H}\| \right]$. In this paper, we use MPC with a window size of 4 hours as a control prior to minimize the risk. MPC $-\epsilon$ is MPC with a generated prediction error of ϵ .

- **Model Predictive Control with LSTM (MPC-LSTM):** Due to the powerful time series prediction ability, Long Short-Term Memory (LSTM) has been utilized as a prediction model in MPC in recent studies [47, 50, 102]. In the water supply control problem, we implement a LSTM model as the predictor and apply it in MPC, which is called MPC-LSTM. The LSTM model has one LSTM layer with 60 hidden neurons and can predict the demand in the future 4 hours. The same training dataset for ML is used for LSTM training.

- **Tube-based Model Predictive Control (TMPC):** TMPC [67, 80] is a computationally efficient robust MPC approach which creates state constraints (tube) based on a nominal dynamic model. TMPC makes sure that the true state of MPC stays within the tube. Since the nominal states are assumed to satisfy the constraint, TMPC can also guarantee a constraint. In our experiments, we design a tube based on a nominal dynamic model exploiting the expected demand information. On the basis of existing TMPC [67, 80], we utilize the LSTM predictor in deciding an action while guaranteeing the action stays in the created tube.

- **Machine Learning (ML):** This is the purely-trained ML model without safety constraints for any episode. For fair comparison, we use the same neural architecture for pure ML and LAOC.

- **Constrained Reinforcement Learning (CRL):** As an important safe reinforcement learning algorithm, CRL [37, 73] has been applied for control problems. Most CRL methods guarantee a constraint in expectation or with a high probability. In our experiments, we implement CRL to satisfy the expected safety risk constraint $\mathbb{E}[R_h^\pi - (1 + \lambda)R_h^{\pi^\dagger}] \leq 0$ with $\lambda = 0$. Not like original model-free CRL, we exploit the dynamic model information for value estimation in RL.

- **Linear Combination (Lin):** Lin- ρ is the policy in Proposition 4.1 that linearly combines ML advice $\tilde{\pi}$ and ROBD π^\dagger as $\pi = \rho\tilde{\pi} + (1 - \rho)\pi^\dagger$ with a combination factor $\rho \in [0, 1]$.

5.2 Results

We provide our main results for the default setting in Table 1. We give the average energy cost, the average carbon cost, and the maximum risk ratio for the control priors, the pure ML model and the learning-augmented algorithms including Lin and LAOC. The results are evaluated for 20 buildings in one year.

First, we can find that the control prior ROBD achieves the lowest safety risk which requires an accurate one-step prediction of the water demand. The control prior OGD which does not rely on any prediction also achieves relatively low risk. However, the average energy costs and carbon emission are relatively large. Assuming a nearly-accurate predictor with a prediction error of 0.03, MPC-0.03 can achieve a maximum risk ratio of 2.52, an average carbon cost of 17782 kg, and an average energy cost of 6924 \$. Thus, MPC has good risk and cost performances when a nearly-accurate predictor is applied. However, a real predictor such as LSTM in our experiments can have large prediction error. The LSTM in our experiment has an average prediction error of 0.05. This results in a higher safety risk and larger average energy cost and carbon emission as is shown by the performance of MPC-LSTM in Table 1. This shows that the performance of MPC is largely affected by the quality of the predictor. Furthermore, we can observe from Table 1 that as a robust MPC method, TMPC can effectively reduce the safety risk, but it has much higher average energy costs and carbon emission.

Different from control priors, the pure ML policy has the lowest energy cost and carbon emission, but it has much higher safety risk ratio. This is because the ML models are trained to optimize the average performances, but they can have arbitrarily bad performance when the adversarial instances exist. With the expected safety constraints, CRL can reduce the safety risk while sacrificing some average cost performance. However, we can observe from Table 1 that the worst-case risk of CRL can still be very high, which is due to the existence of adversarial instances. The vulnerability of ML and CRL impedes their deployments in real water supply systems which are critical for the safety of the buildings.

The learning-augmented designs are given to achieve a tradeoff between the average performance and the worst-case risk. As a naive learning-augmented design, Lin can reduce the safety risk to some extent by choosing a proper combination weight ρ . However, we can find that if we choose a large weight for ML model (e.g. $\rho = 0.5$ in Table 1), the average performance can be good but the maximum risk ratio is still very high. Actually, by Lin-0.5, the

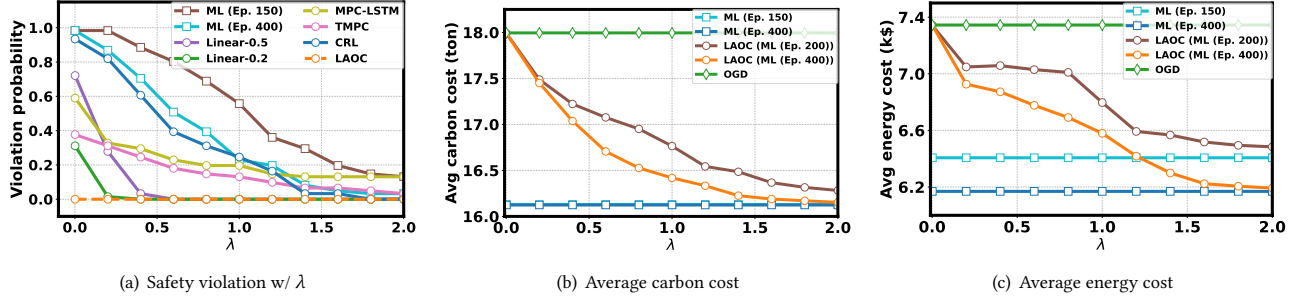


Figure 2: Safety constraint violation and average costs. By default, OGD is the control prior for LAOC. ML (Ep. N) is the ML model at the N th epoch. LAOC (ML (Ep. N)) is LAOC using the purely-trained ML model at the N th epoch.

$(1 + \lambda)$ -safety constraint is violated with a high probability as we will show in Figure 2(a). If we set a small weight for ML model (e.g. $\rho = 0.2$ in Table 1), we can get a low safety risk ratio, but the average costs becomes very large. LAOC is designed to optimizing the average performance while guaranteeing the $(1 + \lambda)$ -safety constraint in (4). We can observe that with a higher safety requirement (e.g. $\lambda = 0.4$), the safety risk ratio is low and close to that of control priors while the average costs are much lower than those of Lin. Also, with a lower safety requirement (e.g. $\lambda = 0.8$), the average costs of LAOC are low and close to those of pure ML, and the risk ratio is also much lower than pure ML because the $(1 + \lambda)$ -safety constraint is always satisfied. Next, we provide more details as below.

5.2.1 Safety Violation. The safety violation probability on testing dataset is given in Figure 2(a). The violation probability is the ratio of the number of safety constraint violation instances to the total testing instance number. A higher λ in $(1 + \lambda)$ -safety in (4) gives a less strict safety constraint, so the violation probability decreases with λ . We can observe that ML can have a high safety violation probability even when λ is large. If the ML model is not sufficiently trained (e.g. ML model at Epoch 150 in Figure 2(a)), the safety violation probability is even higher. These show that pure ML is not safe enough for water supply systems. Although CRL reduces the safety violation probability comparing to ML, it still has a high safety violation rate. Moreover, MPC-LSTM has a high safety violation rates due to the lack of prediction performance guarantee, and TMPC has a reduced but non-zero safety violation probability. As a learning-augmented design, Lin can also violates safety constraint especially when the safety requirement is high (small λ). Decreasing the combination weight for ML model from 0.5 to 0.2 can reduce the violation probability, but this results in a large increase of average costs shown in Table 1. By contrast, LAOC never violates safety constraint given any problem instance and any safety requirement parameter λ , which validates the effectiveness of LAOC in strictly guaranteeing the safety constraint as proved in Theorem 4.3.

5.2.2 Cost-safety tradeoff. Figure 2(b) and Figure 2(c) demonstrates the tradeoff between the average costs and safety requirement for LAOC. The preset parameter λ in the safety constraint (4) indicates the level of safety requirement: with smaller λ , the safety constraint becomes more strict. When $\lambda = 0$, the safety constraint is so strict

that LAOC reduces to the control prior OGD which is the default control prior used in LAOC. Thus, the carbon and energy costs of LAOC is the same as those of OGD. When λ becomes larger, $(1 + \lambda)$ -safety constraint (4) becomes less strict, the average costs of LAOC approaches the average costs of corresponding pure ML models, so LAOC can achieve less carbon and energy costs. When λ becomes large enough, we can observe that the average costs of LAOC are the same as those of pure ML model. The carbon and energy costs also show the impacts of the ML quality. The ML model at Epoch 400 is better than the ML model at Epoch 200, so the average costs of LAOC with ML model at Epoch 400 are lower than those of LAOC with ML model at Epoch 200. These observations coincide with Theorem 4.4 which theoretically shows the tradeoff between the average costs and safety requirement.

5.2.3 OOD Testing. To further evaluate the robustness of the algorithms, we give results under the Out of Distribution (OOD) setting in Table 2. We generate the OOD testing demand sequences by adding Gaussian noise to the original demand sequences. We can find that the control priors ROBD and OGD both achieve low enough safety risk ratio, but their average costs are very high. The LSTM predictor is largely affected by the OOD testing, causing a very high safety risk for MPC-LSTM. TMPC can be applied to reduce the safety risk to some extent, but the worst-case safety risk is still very high. This is because the nominal model in TMPC cannot define a robust enough tube in OOD setting.

The pure ML policy and CRL policy are also largely affected by OOD testing. We can observe from Table 2 that ML and CRL both have high safety risk in the worst case. The expected safety constraint satisfaction in CRL does not help a lot in OOD testing because CRL is trained on a distribution that is very different from the testing distribution. That being said, ML still achieves the lowest average energy and carbon costs.

The learning-augmented designs that combine ML with control priors can take an effect in achieving a low enough safety risk. Even Lin can achieve a low safety risk by choosing a good combination weight ρ . However, Lin has high average energy and carbon costs because it is limited in exploiting the ML predictions. Comparably, LAOC (e.g. $\lambda = 0.4$) not only guarantees a small enough risk for any problem instance, but also achieves low average energy and carbon costs.

6 Concluding Remarks

This work considers an online control problem for water supply management. Besides minimizing the average energy cost, we consider the safety constraint against a given control prior. We design a learning-augmented algorithm, LAOC, that strictly ensure safety constraint. Our analysis reveals the tradeoff between the cost performance and the safety requirement. We evaluate the performance for a case study of building water supply, showing the superiority of LAOC in reducing energy cost and carbon emission and guaranteeing the safety requirements. In the future, the proposed design can be extended to broader applications such as EV charging and sustainable data centers to improve the efficiency and provide safety guarantee for these systems.

Acknowledgments

Pengfei Li and Shaolei Ren were supported by NSF grants CCF-2324916 and CCF-2324941. Adam Wierman was supported by NSF grants CCF-2326609, CNS-2146814, CPS-2136197, CNS-2106403, and NGSDI-2105648 and the funding from the Resnick Sustainability Institute.

References

- [1] 2023. Planning for Sustainable Water Infrastructure. <https://www.epa.gov/sustainable-water-infrastructure/planning-sustainable-water-infrastructure>.
- [2] 2024. Energy Efficiency for Water Utilities. <https://www.epa.gov/sustainable-water-infrastructure/energy-efficiency-water-utilities>.
- [3] 2024. Exploring the interdependence of two critical resources Energy and Water. <https://www.iea.org/topics/energy-and-water>.
- [4] 2024. Horsepower required to pump water. https://www.engineeringtoolbox.com/pumping-water-horsepower-d_753.html.
- [5] 2024. Real-time Price by CAISO. http://www.energyonline.com/Data/GenericData.aspx?DataId=19&CAISO__Real-time_Price/.
- [6] 2024. STRATEGIES FOR SAVING ENERGY AT PUBLIC WATER SYSTEMS. <https://www.epa.gov/sites/default/files/2015-04/documents/epa816f13004.pdf>.
- [7] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *International conference on machine learning*. PMLR, 22–31.
- [8] Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. 2019. Differentiable convex optimization layers. *Advances in neural information processing systems* 32 (2019).
- [9] Mohammad Ali Alomrani, Reza Moravej, and Elias Boutros Khalil. 2022. Deep Policies for Online Bipartite Matching: A Reinforcement Learning Approach. *Transactions on Machine Learning Research* (2022). <https://openreview.net/forum?id=mbwm7NdkpO>
- [10] Sanae Amani, Christos Thrampoulidis, and Lin Yang. 2021. Safe Reinforcement Learning with Linear Function Approximation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 243–253. <https://proceedings.mlr.press/v139/amani21a.html>
- [11] BD Anderson and John B Moore. 2012. Optimal Filtering. Courier Corporation. *Courier Corporation* (2012).
- [12] Antonios Antoniadis, Christian Coester, Marek Elias, Adam Polak, and Bertrand Simon. 2020. Online Metric Algorithms with Untrusted Predictions. In *ICML*.
- [13] 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.
- [14] Gissella Bejarano, Adita Kulkarni, Raushan Raushan, Anand Seetharam, and Arti Ramesh. 2019. Swap: Probabilistic graphical and deep learning models for water consumption prediction. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 233–242.
- [15] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. 2004. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised Lectures* (2004), 169–207.
- [16] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. 2022. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems* 5, 1 (2022), 411–444.
- [17] Eduardo F Camacho and Carlos Bordons Alba. 2013. *Model predictive control*. Springer.
- [18] Agustín Castellano, Hancheng Min, Juan Bazerque, and Enrique Mallada. 2022. Reinforcement Learning with Almost Sure Constraints. In *Learning for Dynamics and Control*.
- [19] Carmen Cheh, Justin Albrethsen, Zhen Wei Ng, Binbin Chen, Xin Lou, Zaki Masood, and David KY Yau. 2024. Water Pump Operation Optimization under Dynamic Market and Consumer Behaviour. (2024), 335–346.
- [20] 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*. 199–210.
- [21] Yuri Chervonyi, Praneet Dutta, Piotr Trochim, Octavian Voicu, Cosmin Padurararu, Crystal Qian, Emre Karagozler, Jared Quincy Davis, Richard Chippendale, Gautam Bajaj, et al. 2022. Semi-analytical industrial cooling system model for reinforcement learning. *arXiv preprint arXiv:2207.13131* (2022).
- [22] Yuri Chervonyi, Praneet Dutta, Piotr Trochim, Octavian Voicu, Cosmin Padurararu, Crystal Qian, Emre Karagozler, Jared Quincy Davis, Richard Chippendale, Gautam Bajaj, et al. 2022. Semi-analytical industrial cooling system model for reinforcement learning. *arXiv preprint arXiv:2207.13131* (2022).
- [23] Nicolas Christianson, Junxuan Shen, and Adam Wierman. 2023. Optimal robustness-consistency tradeoffs for learning-augmented metrical task systems. In *AI STATS*.
- [24] Joshua Comden, Sijie Yao, Niangjun Chen, Haipeng Xing, and Zhenhua Liu. 2019. Online Optimization in Cloud Resource Provisioning: Predictions, Regrets, and Algorithms. *Proc. ACM Meas. Anal. Comput. Syst.* 3, 1, Article 16 (March 2019), 30 pages. <https://doi.org/10.1145/3322205.3311087>
- [25] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. 2021. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3304–3312.
- [26] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. 2020. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems* 33 (2020), 8378–8390.
- [27] Bingqian Du, Chuan Wu, and Zhiyi Huang. 2019. Learning Resource Allocation and Pricing for Cloud Profit Maximization. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) (AAAI'19/AAAI'19/EAAI'19). AAAI Press, Article 929, 8 pages. <https://doi.org/10.1609/aaai.v33i01.33017570>
- [28] Claudia Dâ€™Ambrosio, Andrea Lodi, Sven Wiese, and Cristiana Bragalli. 2015. Mathematical programming techniques in water network optimization. *European Journal of Operational Research* 243, 3 (2015), 774–788.
- [29] Yonathan Efroni, Shie Mannor, and Matteo Pirodda. 2020. Exploration-exploitation in constrained mdp. *arXiv preprint arXiv:2003.02189* (2020).
- [30] David D Fan, Ali-akbar Agha-mohammadi, and Evangelos A Theodorou. 2020. Deep learning tubes for tube mpc. *arXiv preprint arXiv:2002.01587* (2020).
- [31] Ana Lúcia D Franco, Henri Bourlés, Edson R De Pieri, and Herve Guillard. 2006. Robust nonlinear control associating robust feedback linearization and H/sub spl infin//control. *IEEE transactions on automatic control* 51, 7 (2006), 1200–1207.
- [32] Randy Freeman and Petar V Kokotovic. 2008. *Robust nonlinear control design: state-space and Lyapunov techniques*. Springer Science & Business Media.
- [33] Aditya Gahlawat, Pan Zhao, Andrew Patterson, Naira Hovakimyan, and Evangelos Theodorou. 2020. L1-GP: L1 adaptive control with Bayesian learning. In *Learning for dynamics and control*. PMLR, 826–837.
- [34] Bissan Ghaddar, Joe Naoum-Sawaya, Akihiro Kishimoto, Nicole Taheri, and Bradley Eck. 2015. A Lagrangian decomposition approach for the pump scheduling problem in water networks. *European Journal of Operational Research* 241, 2 (2015), 490–501.
- [35] Zabih Ghelichi, Javad Tajik, and Mir Saman Pishvae. 2018. A novel robust optimization approach for an integrated municipal water distribution system design under uncertainty: A case study of Mashhad. *Computers & Chemical Engineering* 110 (2018), 13–34.
- [36] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. 2022. Provably efficient model-free constrained rl with linear function approximation. *arXiv preprint arXiv:2206.11889* (2022).
- [37] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. 2022. Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems* 35 (2022), 13303–13315.
- [38] Gautam Goel, Naman Agarwal, Karan Singh, and Elad Hazan. 2022. Best of Both Worlds in Online Control: Competitive Ratio and Policy Regret. *arXiv preprint arXiv:2211.11219* (2022).
- [39] Gautam Goel and Babak Hassibi. 2021. Competitive Control. <https://doi.org/10.48550/ARXIV.2107.13657>

- [40] Gautam Goel and Babak Hassibi. 2022. Competitive control. *IEEE Trans. Automat. Control* (2022).
- [41] Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. 2019. Beyond Online Balanced Descent: An Optimal Algorithm for Smoothed Online Optimization. In *NeurIPS*, Vol. 32. <https://proceedings.neurips.cc/paper/2019/file/9f36407ead0629fc166f14dde7970f68-Paper.pdf>
- [42] Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. 2019. Beyond online balanced descent: an optimal algorithm for smoothed online optimization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 168, 11 pages.
- [43] Gautam Goel and Adam Wierman. 2019. An Online Algorithm for Smoothed Online Convex Optimization. *SIGMETRICS Perform. Eval. Rev.* 47, 2 (Dec. 2019), 646–68.
- [44] Alexander P Goryashko and Arkadi S Nemirovski. 2014. Robust energy cost optimization of water distribution system with uncertain demand. *Automation and Remote Control* 75 (2014), 1754–1769.
- [45] Moritz Hardt and Max Simchowitz. 2018. Convex Optimization and Approximation. <https://ee227c.github.io/notes/ee227c-notes.pdf>.
- [46] Lukas Hewing, Kim P Wabersich, Marcel Menner, and Melanie N Zeilinger. 2020. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems* 3, 1 (2020), 269–296.
- [47] Keke Huang, Ke Wei, Fanbiao Li, Chunhua Yang, and Weihua Gui. 2022. LSTM-MPC: A deep learning based predictive control method for multimode process control. *IEEE Transactions on Industrial Electronics* 70, 11 (2022), 11544–11554.
- [48] Girish Joshi and Girish Chowdhary. 2019. Deep model reference adaptive control. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 4601–4608.
- [49] Girish Joshi, Jasvir Virdi, and Girish Chowdhary. 2021. Asynchronous deep model reference adaptive control. In *Conference on Robot Learning*. PMLR, 984–1000.
- [50] Marvin Jung, Paulo Renato da Costa Mendes, Magnus Önnheim, and Emil Gustavsson. 2023. Model Predictive Control when utilizing LSTM as dynamic models. *Engineering Applications of Artificial Intelligence* 123 (2023), 106226.
- [51] IS Khalil, JC Doyle, and K Glover. 1996. *Robust and optimal control*. Prentice hall.
- [52] Ridiger Kiesel and Michael Kusterman. 2016. Structural models for coupled electricity markets. *Journal of Commodity Markets* 3, 1 (2016), 16–38.
- [53] Zachary J Lee, Tongxin Li, and Steven H Low. 2019. ACN-data: Analysis and applications of an open EV charging dataset. In *Proceedings of the tenth ACM international conference on future energy systems*. 139–149.
- [54] Pengfei Li, Jianyi Yang, and Shaolei Ren. 2022. Expert-Calibrated Learning for Online Optimization with Switching Costs. In *SIGMETRICS*.
- [55] Pengfei Li, Jianyi Yang, and Shaolei Ren. 2022. Expert-Calibrated Learning for Online Optimization with Switching Costs. *Proc. ACM Meas. Anal. Comput. Syst.* 6, 2, Article 28 (Jun 2022), 35 pages.
- [56] Pengfei Li, Jianyi Yang, and Shaolei Ren. 2023. Learning for Edge-Weighted Online Bipartite Matching with Robustness Guarantees. *ICML* (2023).
- [57] Pengfei Li, Jianyi Yang, and Shaolei Ren. 2023. Robustified Learning for Online Optimization with Memory Costs. *INDOCOM* (2023).
- [58] Tongxin Li, Bo Sun, Yue Chen, Zixin Ye, Steven H Low, and Adam Wierman. 2021. Learning-based Predictive Control via Real-time Aggregate Flexibility. *IEEE Transactions on Smart Grid* 12, 6 (2021), 4897–4913.
- [59] Tongxin Li, Ruixiao Yang, Guannan Qu, Guanya Shi, Chenkai Yu, Adam Wierman, and Steven Low. 2022. Robustness and Consistency in Linear Quadratic Control with Untrusted Predictions. *Proc. ACM Meas. Anal. Comput. Syst.* 6, 1, Article 18 (feb 2022), 35 pages. <https://doi.org/10.1145/3508038>
- [60] Yingying Li, Xin Chen, and Na Li. 2019. Online optimal control with linear dynamics and predictions: Algorithms and regret analysis. *Advances in Neural Information Processing Systems* 32 (2019).
- [61] Yingying Li and Na Li. 2020. Leveraging Predictions in Smoothed Online Convex Optimization via Gradient-based Algorithms. In *NeurIPS*, Vol. 33. <https://proceedings.neurips.cc/paper/2020/file/a6e4f250fb5c56aaf215a236c64e5b0a-Paper.pdf>
- [62] Yingying Li, Jing Yu, Lauren Conger, Taylan Kargin, and Adam Wierman. [n. d.]. Learning the Uncertainty Sets of Linear Control Systems via Set Membership: A Non-asymptotic Analysis. In *Forty-first International Conference on Machine Learning*.
- [63] Enming Liang, Minghua Chen, and Steven H. Low. 2023. Low Complexity Homeomorphic Projection to Ensure Neural-Network Solution Feasibility for Optimization over (Non-)Convex Set. In *ICML*.
- [64] Yiheng Lin, Judy Gan, Guannan Qu, Yash Kanoria, and Adam Wierman. 2022. Decentralized Online Convex Optimization in Networked Systems. In *International Conference on Machine Learning*. PMLR, 13356–13393.
- [65] Yiheng Lin, Yang Hu, Guanya Shi, Haoyuan Sun, Guannan Qu, and Adam Wierman. 2021. Perturbation-based Regret Analysis of Predictive Control in Linear Time Varying Systems. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 5174–5185. <https://proceedings.neurips.cc/paper/2021/file/298f587406c914fad5373bb689300433-Paper.pdf>
- [66] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. 2021. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems* 34 (2021), 17183–17193.
- [67] Brett T Lopez, Jean-Jacques E Slotine, and Jonathan P How. 2019. Dynamic tube MPC for nonlinear systems. In *2019 American Control Conference (ACC)*. IEEE, 1655–1662.
- [68] Tiago Luna, João Ribau, David Figueiredo, and Rita Alves. 2019. Improving energy efficiency in water supply systems with pump scheduling optimization. *Journal of cleaner production* 213 (2019), 342–356.
- [69] Jerry Luo, Cosmin Paduraru, Octavian Voicu, Yuri Chervonyi, Scott Munns, Jerry Li, Crystal Qian, Praneet Dutta, Jared Quincy Davis, Ningjia Wu, et al. 2022. Controlling Commercial Cooling Systems Using Reinforcement Learning. *arXiv preprint arXiv:2211.07357* (2022).
- [70] Electricity Maps. 2024. Carbon Intensity Data (Version January 17, 2024). *Electricity Maps Data Portal* (2024). <https://www.electricitymaps.com/data-portal>
- [71] Konstantinos Oikonomou and Masood Parvania. 2018. Optimal coordination of water distribution energy flexibility with power systems operation. *IEEE Transactions on Smart Grid* 10, 1 (2018), 1101–1110.
- [72] Weici Pan, Guanya Shi, Yiheng Lin, and Adam Wierman. 2022. Online Optimization with Feedback Delay and Nonlinear Switching Cost. *Proc. ACM Meas. Anal. Comput. Syst.* 6, 1, Article 17 (Feb 2022), 34 pages. <https://doi.org/10.1145/3508037>
- [73] Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. 2019. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems* 32 (2019).
- [74] Marios M Polycarpou and Petros A Ioannou. 1993. A robust adaptive nonlinear control design. In *1993 American control conference*. IEEE, 1365–1369.
- [75] Claudia Quintiliani and Enrico Creaco. 2019. Using additional time slots for improving pump control optimization based on trigger levels. *Water Resources Management* 33 (2019), 3175–3186.
- [76] Daan Rutten, Nico Christianson, Debankur Mukherjee, and Adam Wierman. 2022. Online Optimization with Untrusted Predictions. *arXiv preprint arXiv:2202.03519* (2022).
- [77] Mohammad A Salahuddin, Ala Al-Fuqaha, and Mohsen Guizani. 2016. Reinforcement learning for resource provisioning in the vehicular cloud. *IEEE Wireless Communications* 23, 4 (2016), 128–135.
- [78] Guanya Shi. 2021. Competitive Control via Online Optimization with Memory, Delayed Feedback, and Inexact Predictions. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*.
- [79] Guanya Shi, Yiheng Lin, Soon-Jo Chung, Yisong Yue, and Adam Wierman. 2020. Online optimization with memory and competitive control. *Advances in Neural Information Processing Systems* 33 (2020), 20636–20647.
- [80] Jerome Sieber, Samir Bennani, and Melanie N Zeilinger. 2021. A system level approach to tube-based model predictive control. *IEEE Control Systems Letters* 6 (2021), 776–781.
- [81] Manish K Singh and Vassilis Kekatos. 2019. Optimal scheduling of water distribution systems. *IEEE Transactions on Control of Network Systems* 7, 2 (2019), 711–723.
- [82] Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyang Wang, David H Mguni, Jun Wang, and Haitham Ammar. 2022. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*. PMLR, 20423–20443.
- [83] Pantelis Sotasakis, Ajay K Sampathirao, Alberto Bemporad, and Panagiotis Patrinos. 2018. Uncertainty-aware demand management of water distribution networks in deregulated energy markets. *Environmental modelling & software* 101 (2018), 10–22.
- [84] Anna Stuhlmacher and Johanna L Mathieu. 2020. Chance-constrained water pumping to manage water and power demand uncertainty in distribution networks. *Proc. IEEE* 108, 9 (2020), 1640–1655.
- [85] Anna Stuhlmacher and Johanna L Mathieu. 2020. Water distribution networks as flexible loads: A chance-constrained programming approach. *Electric Power Systems Research* 188 (2020), 106570.
- [86] Chenxi Sun, Tongxin Li, and Xiaoying Tang. 2021. Data-driven Electric Vehicle Charging Station Placement for Incentivizing Potential Demand. In *2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 27–32.
- [87] Wei Sun and Chenchen Huang. 2022. Predictions of carbon emission intensity based on factor analysis and an improved extreme learning machine from the perspective of carbon emission efficiency. *Journal of Cleaner Production* 338 (2022), 130414.
- [88] Wentao Tang and Prodromos Daoutidis. 2022. Data-driven control: Overview and perspectives. In *2022 American Control Conference (ACC)*. IEEE, 1048–1064.
- [89] Andrew Taylor, Andrew Singletary, Yisong Yue, and Aaron Ames. 2020. Learning for safety-critical control with control barrier functions. In *Learning for Dynamics and Control*. PMLR, 708–717.

- [90] Vestal Tutterow and Aimee T McKane. 2004. Variable speed pumping: A guide to successful applications. US DOE. Lawrence Berkeley National Laboratory. <https://escholarship.org/uc/item/4691d71q>.
- [91] Michael Volk. 2013. *Pump characteristics and applications*. CRC Press.
- [92] Kim P Wabersich, Lukas Hewing, Andrea Carron, and Melanie N Zeilinger. 2021. Probabilistic model predictive safety certification for learning-based control. *IEEE Trans. Automat. Control* 67, 1 (2021), 176–188.
- [93] Ye Wang, Kevin Too Yok, Wenyan Wu, Angus R Simpson, Erik Weyer, and Chris Manzie. 2021. Minimizing pumping energy cost in real-time operations of water distribution systems using economic model predictive control. *Journal of Water Resources Planning and Management* 147, 7 (2021), 04021042.
- [94] Evan M Wanjiru, Lijun Zhang, and Xiaohua Xia. 2016. Model predictive control strategy of energy-water management in urban households. *Applied Energy* 179 (2016), 821–831.
- [95] Honghao Wei, Xin Liu, and Lei Ying. 2021. A provably-efficient model-free algorithm for constrained markov decision processes. *arXiv preprint arXiv:2106.01577* (2021).
- [96] William Wong, Praneet Dutta, Octavian Voicu, Yuri Chervonyi, Cosmin Padurararu, and Jerry Luo. 2022. Optimizing industrial hvac systems with hierarchical reinforcement learning. *arXiv preprint arXiv:2209.08112* (2022).
- [97] Jianyi Yang and Shaolei Ren. 2022. Learning-Assisted Algorithm Unrolling for Online Optimization with Budget Constraints. *AAAI* (2022).
- [98] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J. Ramadge. 2020. Projection-Based Constrained Policy Optimization. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rke3TJrtPS>
- [99] Yunchang Yang, Tianhao Wu, Han Zhong, Evrard Garcelon, Matteo Pirotta, Alessandro Lazaric, Liwei Wang, and Simon Shaolei Du. 2022. A Reduction-Based Framework for Conservative Bandits and Reinforcement Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=AclrgZ9BKed>
- [100] Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, and Adam Wierman. 2020. The power of predictions in online control. *Advances in Neural Information Processing Systems* 33 (2020), 1994–2004.
- [101] Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, and Adam Wierman. 2022. Competitive control with delayed imperfect information. In *2022 American Control Conference (ACC)*. IEEE, 2604–2610.
- [102] Krzysztof Zarzycki and Maciej Ławryńczuk. 2024. LSTM for Modelling and Predictive Control of Multivariable Processes. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 74–87.
- [103] Runyu Zhang, Yingying Li, and Na Li. 2021. On the regret analysis of online LQR control with predictions. In *2021 American Control Conference (ACC)*. IEEE, 697–703.

Appendix

A Additional Numerical Results

In this section, we give more numerical results of the case study in Section 5. We first give more details on the testing loss for different training epochs and safety parameter λ followed by the maximum risk ratio with different λ . Then, we provide an ablation study on the impact of different control priors on LAOC. Next, we show the safety violation probability under the OOD setting. Finally, we give an instance study to explain LAOC intuitively.

A.1 Training and Testing Details

A.1.1 Convergence. In Figure 3(a), we show the average testing losses as the training evolves. We show the training sequences of pure ML (blue curve) and the safety-aware fine-tuning of LAOC-F (orange curve), respectively. LAOC (ML) (green curve) takes the purely trained ML model at the corresponding epoch as input. The average loss is normalized by the average loss of the optimal policy, i.e. $\mathbb{E}[J_H^z]/\mathbb{E}[J_H^*]$. The testing losses converge after 400 epochs. We can find that ML purely trained without considering safety has the best testing loss convergence. Due to the safety constraint, the testing loss of LAOC (ML) with the purely-trained ML model as input increases a lot. By the safety-aware finetuning in (10), LAOC-F effectively reduces the testing loss of LAOC because the safety-aware finetuning is performed on an objective that takes the safety set (7) into consideration, which validates the conclusion in Theorem 4.5.

A.1.2 Testing loss with respect to λ . Figure 3(b) shows the the average testing loss changing with the safety parameter λ in the safety constraint (4) for LAOC. The average testing loss is the weighted combination of the energy cost, carbon cost and the deviation penalty, and is normalized by the average loss of the optimal policy. When λ becomes larger, $(1 + \lambda)$ -safety constraint (4) becomes less strict, the average loss of LAOC approaches the average loss of corresponding purely-trained ML. When $\lambda = 0$, the safety constraint is the strictest and LAOC reduces to the control prior OGD. These observations coincide with the cost bound in Theorem 4.4. Additionally, we evaluate the average testing loss of Lin-0.2 and find that although Lin-0.2 has low safety violation probability, it is so conservative that average loss is very high. These validate the superiority of LAOC in achieving a low enough average loss while guaranteeing safety.

A.1.3 Maximum risk ratio with respect to λ . In Figure 3(c), we show the worst-case risk ratio changing with the safety parameter λ in the safety constraint (4). If the safety requirement parameter λ becomes larger, LAOC will take greater risks. Nevertheless, the risk is still lower than purely-trained ML even with very large λ . These results show the advantage of LAOC in decarbonizing the water supply systems under the safety guarantee.

A.2 LAOC with different control priors

In Figure 4, we give the average costs of LAOC using different control priors (OGD,ROBD,MPC). Here, MPC represents MPC-0.03 with a generated prediction error of 0.03. MPC-0.03 can achieve a maximum risk ratio of 2.52, an average carbon cost of 17782 kg, and an average energy cost of 6924 \$. By the performance bound in Theorem 4.4, the expected loss is affected by the per-round risk performance of the control prior r_h^\dagger and the action discrepancy δ_h between the

pure ML action and the control prior. As shown in Table 1, ROBD has the lowest worst-case risk which defines the most stringent safety constraint, so LAOC (ROBD) has larger average loss and larger carbon/energy costs than LAOC with the other two priors. We also observe that although OGD has the largest average carbon/energy costs, LAOC (OGD) can achieve low carbon/energy costs a when λ is slightly larger. This is because the safety constraint is defined by the risk of OGD which is higher than that of ROBD. No matter which control prior is considered, LAOC can always guarantee the $(1 + \lambda)$ -safety constraint with respect to the control prior.

A.3 Safety Violation Probability Under OOD Setting

Under the OOD setting, the violation rates of safety constraint (4) with respect to the control prior OGD are given in Figure 5. A higher λ in $(1 + \lambda)$ -safety in (4) gives a less strict safety constraint, so the violation probability decreases with λ . We can observe that MPC-LSTM is largely affected by the distribution shift and has the highest safety violation probability. TMPC reduces the violation probability but still has a large violation probability. Both ML and CRL have non-zero violation probability. We can find that the violation probability of CRL is even larger than the violation probability of ML when λ is small. The ineffectiveness of CRL is because CRL guarantees an expected constraint on the training distribution but the testing distribution has been very different from the training distribution.

As a learning-augmented design, Lin can achieve low safety constraint violation rate by choosing a small enough combination weight, but this results in a large increase of average costs shown in Table 2. By contrast, even in the OOD setting, LAOC never violates safety constraint given any problem instance and any safety requirement parameter λ , which validates the effectiveness of LAOC in strictly guaranteeing the safety constraint as proved in Theorem 4.3.

A.4 Instance study

In Figure 6, we give a snapshot of a problem instance with 24 hours to get better intuitions on the control process. Figure 6(a) shows the traces of carbon intensity, energy price, and water demand of the instance. From Figure 6(b), we can observe that ML chooses to delay the water supply when the carbon intensity or energy price is high. ML tends to schedule a large water supply when the carbon intensity or energy price is relatively low. This shows the effectiveness of ML policy in utilizing the water tank to save the energy costs by buffering the demand. However, from Figure 6(c), we can find that the water level of ML can be very low at some hour. In this instance, the water level by ML can reduce to $10 m^3$ which is much lower than the nominal safe water level $\bar{x} = 40m^3$. This results in a high safety risk since the water is not enough when there is an emergency in the building. Comparably, the control priors OGD and ROBD take much more conservative action shown in Figure 6(b) and maintain the nominal water level very well shown in Figure 6(c). However, they are limited in predicting and exploiting the time-varying energy price and carbon intensity, thus ineffective in saving energy costs and reducing carbon emissions. Different from them, the proposed algorithm LAOC ($\lambda = 0.8$) achieves a good

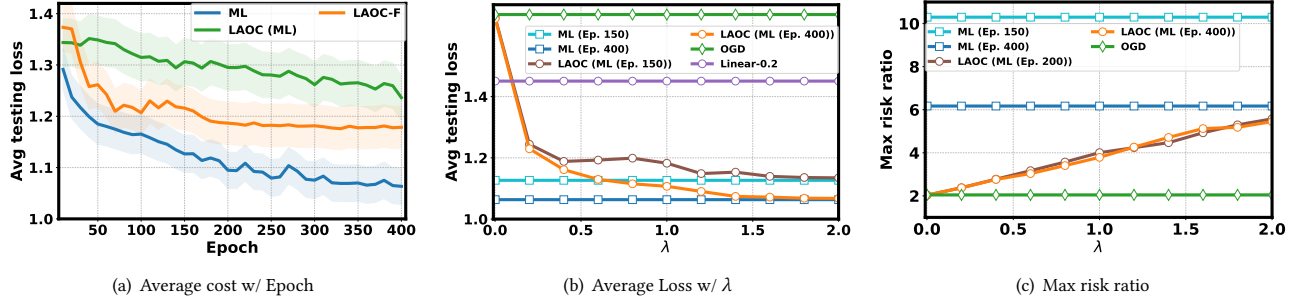


Figure 3: Average testing loss and the maximum risk ratio. By default, OGD is the control prior for LAOC. ML (Ep. N) is the ML model at the N th epoch. LAOC (ML (Ep. N)) is LAOC using the purely-trained ML model at the N th epoch. LAOC-F is LAOC with safety-aware finetuning (10).

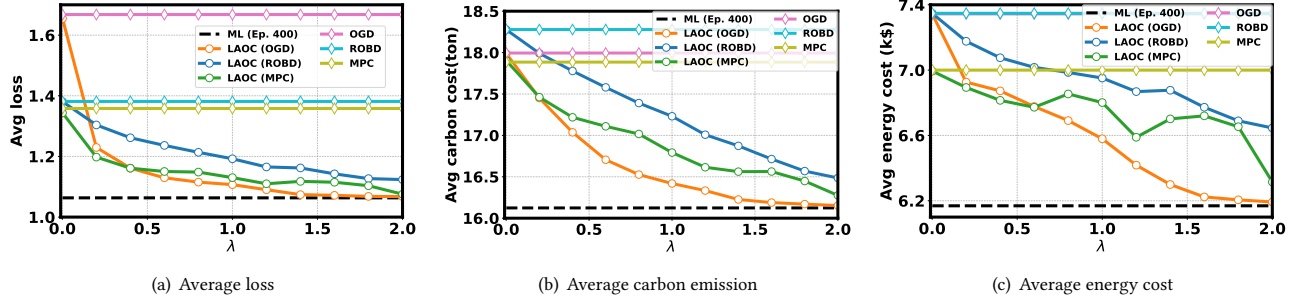


Figure 4: Average loss, carbon cost and energy cost for different LAOC algorithms and control priors. LAOC algorithms use purely-trained ML model at Epoch 400. Here, MPC represents MPC-0.03 with a prediction error of 0.03.

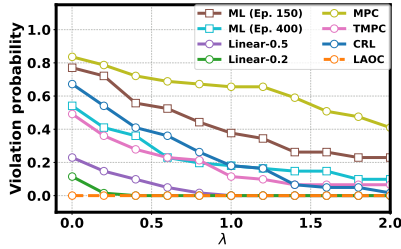


Figure 5: Safety Violation Probability Under OOD Setting.

trade-off between safety and costs. It can maintain a water level not far from nominal water level (orange curve in Figure 6(c)), so the safety risk of LAOC is low. At the same time, LAOC regulates the water supply aware of the time-varying carbon intensity and energy price (orange curve in Figure 6(b)), so it is also effective in saving energy costs and reducing carbon emissions.

B Proof of Proposition 4.1

PROOF. We prove by providing a contradictory example. In this example, the dynamic function is linear, i.e. $f_h(x, u) = \sigma_x x_h + \sigma_u u_h + w_h$, and the control prior has a competitive ratio of η_{π^\dagger} (i.e. $\frac{R_H^\dagger}{R_H^*} \leq \eta_{\pi^\dagger}$). We prove that at least for this example, λ -competitiveness is not guaranteed by Lin.

If Lin guarantees λ -competitiveness, since the competitive ratio of π^\dagger is η_{π^\dagger} , we must have

$$R_H^{\text{Lin}} \leq (1 + \lambda) R_H^{\pi^\dagger} \leq (1 + \lambda) \eta_{\pi^\dagger} R_H^{\pi^*}. \quad (16)$$

Since the risk function r_h is α -strongly convex and the dynamic function is linear, the total risk R_H is also strongly convex with parameter α . By the smoothness of the cost function, we have $\nabla_{u^*} R_H^{\pi^*} = 0$, and so

$$R_H^{\text{Lin}} \geq R_H^{\pi^*} + \frac{\alpha}{2} \|\rho \tilde{u} + (1 - \rho) u^\dagger - u^*\|^2, \quad (17)$$

where the inequality holds by α -strongly convexity of $R_H(u)$. Substituting (17) into (16), we have

$$\frac{\alpha}{2} \|\rho \tilde{u} + (1 - \rho) u^\dagger - u^*\|^2 \leq ((1 + \lambda) \eta_{\pi^\dagger} - 1) R_H^{\pi^*}, \quad (18)$$

and by moving items and the triangle inequality, we have

$$\|\rho(\tilde{u} - u^*)\| - \|(1 - \rho)(u^\dagger - u^*)\| \leq \sqrt{\frac{2}{\alpha} ((1 + \lambda) \eta_{\pi^\dagger} - 1) R_H^{\pi^*}}. \quad (19)$$

Applying the α -strongly convex of $R_H(u)$ and $\nabla_{u^*} R_H^{\pi^*} = 0$ again, we have

$$R_H^{\pi^\dagger} \geq R_H^{\pi^*} + \frac{\alpha}{2} \|u^\dagger - u^*\|^2. \quad (20)$$

Substituting (20) into (19), we have

$$\begin{aligned} \|\tilde{u} - u^*\| &\leq \frac{1 - \rho}{\rho} \sqrt{\frac{2}{\alpha} (R_H^{\pi^\dagger} - R_H^{\pi^*})} + \frac{1}{\rho} \sqrt{\frac{2}{\alpha} ((1 + \lambda) \eta_{\pi^\dagger} - 1) R_H^{\pi^*}} \\ &\leq \sqrt{\frac{2}{\alpha}} \left(\frac{1 - \rho}{\rho} \sqrt{\eta_{\pi^\dagger} - 1} + \frac{1}{\rho} \sqrt{(1 + \lambda) \eta_{\pi^\dagger} - 1} \right) \sqrt{R_H^{\pi^*}} \end{aligned} \quad (21)$$

If Lin guarantees the λ -competitiveness, then the ML advice must satisfy

$$\frac{\|\tilde{u} - u^*\|^2}{R_H^*} \leq \frac{2}{\alpha} \left(\frac{1 - \rho}{\rho} \sqrt{\eta_{\pi^\dagger} - 1} + \frac{1}{\rho} \sqrt{(1 + \lambda) \eta_{\pi^\dagger} - 1} \right)^2. \quad (22)$$

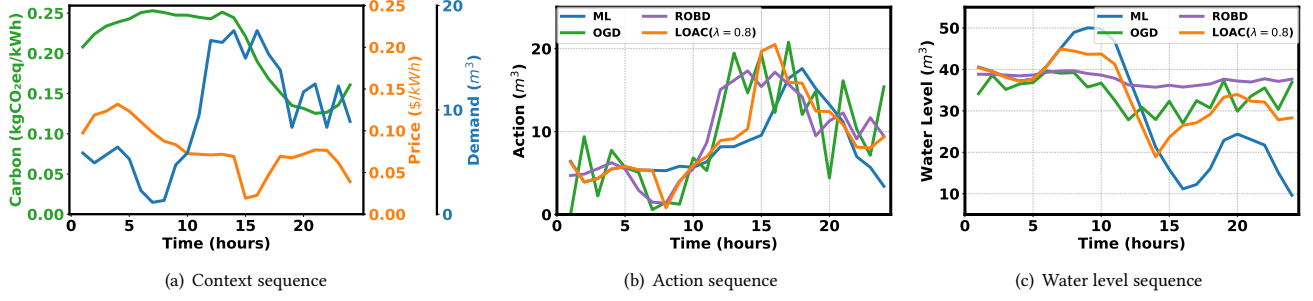


Figure 6: A sequence snapshot of 24 hours. Contexts include carbon intensity, energy price and demand.

Given $\rho \in (0, 1]$ and finite η_{π^\dagger} , the right-hand-side is a finite value. Thus, when $\rho \neq 0$, for arbitrary ML advice with unbounded $\frac{\|\hat{u} - u^*\|^2}{R_H^\dagger}$, λ -competitiveness is not guaranteed. \square

C Proof of Proposition 4.2

Lemma C.1 ([45]). *For any convex and β -cost function r with respect to its input (x, u) , it holds for a parameter $\lambda > 0$ that,*

$$r(x, u) - (1 + \lambda)r(x^\dagger, u^\dagger) \leq (1 + \frac{1}{\lambda}) \left(\frac{\beta}{2} \|x - x^\dagger\|^2 + \frac{\beta}{2} \|u - u^\dagger\|^2 \right) \quad (23)$$

Proof of Proposition 4.2

PROOF. Note that ϕ_h is non-negative. Thus, if the safe action set $\mathcal{U}_{\lambda, h}$ in (7) is non-empty for each $h \in [H]$, then we can always guarantee the competitiveness in Eqn. (5) by selecting an action in $\mathcal{U}_{\lambda, h}$ in Algorithm 1. Then we prove the non-empty of safe action set $\mathcal{U}_{\lambda, h}$ by induction.

First of all, $\mathcal{U}_{\lambda, 0}$ is not empty because u_0^\dagger is always in $\mathcal{U}_{\lambda, 0}$. Then assuming $\mathcal{U}_{\lambda, h-1}$ is not empty, we prove $\mathcal{U}_{\lambda, h}$ is not empty and at least contain an action u_h^\dagger as follows. Since $\mathcal{U}_{\lambda, h-1}$ is not empty, we have $u_{h-1} \in \mathcal{U}_{\lambda, h-1}$ by Algorithm 1, and it holds that

$$R_{h-1} + \phi_{h-1}(u_{h-1}) \leq (1 + \lambda)R_{h-1}^\dagger. \quad (24)$$

Thus if $u_h = u_h^\dagger$, we have

$$\begin{aligned} & R_{h-1} + r_t(x_h, u_h^\dagger) + \phi_h(u_h^\dagger) - (1 + \lambda) \left(R_{h-1}^\dagger + r_h(x_h^\dagger, u_h^\dagger) \right) \\ &= R_{h-1} - (1 + \lambda)R_{h-1}^\dagger + \left(r_h(x_h, u_h^\dagger) - (1 + \lambda)r_h(x_h^\dagger, u_h^\dagger) \right) + \phi_h(u_h^\dagger) \\ &\leq -\phi_{h-1}(u_{h-1}) + \phi_h(u_h^\dagger) + \left(r_h(x_h, u_h^\dagger) - (1 + \lambda)r_h(x_h^\dagger, u_h^\dagger) \right) \\ &\leq -\phi_{h-1}(u_{h-1}) + \phi_h(u_h^\dagger) + (1 + \frac{1}{\lambda}) \frac{\beta}{2} \|x_h - x_h^\dagger\|^2 \end{aligned} \quad (25)$$

where the second inequality holds by Lemma C.1.

Since the reservation cost is chosen as

$$\phi_h(u) = q_h \|f_h(x_h, u) - f_h(x_h^\dagger, u_h^\dagger)\|^2, \quad (26)$$

we have

$$\begin{aligned} & -\phi_{h-1}(u_{h-1}) + \phi_h(u_h^\dagger) \\ &= -q_{h-1} \|x_h - x_h^\dagger\|^2 + q_h \|f_h(x_h, u_h^\dagger) - f_h(x_h^\dagger, u_h^\dagger)\|^2 \\ &\leq (-q_{h-1} + q_h \sigma_x^2) \|x_h - x_h^\dagger\|^2 \\ &\leq - (1 + \frac{1}{\lambda}) \frac{\beta}{2} \|x_h - x_h^\dagger\|^2, \end{aligned} \quad (27)$$

where the first inequality comes from the Lipschitz continuity of dynamic f_h , and the second inequality holds by the choice of $q_h = C_1(1 + \frac{1}{\lambda}) \frac{\beta}{2} \sum_{h'=0}^{H-h-1} (C_2 \sigma_x^2)^{h'}$ for some constant $C_1 \geq 1$ and $C_2 \geq 1$ such that $q_h \sigma_x^2 = C_1(1 + \frac{1}{\lambda}) \frac{\beta}{2} \sum_{h'=1}^{H-h} C_2^{h'-1} \sigma_x^{2h'} \leq C_1(1 + \frac{1}{\lambda}) \frac{\beta}{2} \sum_{h'=1}^{H-h} C_2^{h'} \sigma_x^{2h'} = q_{h-1} - C_1(1 + \frac{1}{\lambda}) \frac{\beta}{2} \leq q_{h-1} - (1 + \frac{1}{\lambda}) \frac{\beta}{2}$.

Substituting (27) into (25), it holds for $u_h = u_h^\dagger$ that $R_{h-1} + r_t(x_h, u_h^\dagger) + \phi_h(u_h^\dagger) \leq (1 + \lambda) \left(R_{h-1}^\dagger + r_h(x_h^\dagger, u_h^\dagger) \right)$. Therefore, u_h^\dagger is in the safe action set $\mathcal{U}_{\lambda, h}$ and so $\mathcal{U}_{\lambda, h}$ is not empty.

Therefore by the discussion at the beginning of this proof, the Proposition is proved. \square

D Proof of Theorem 4.4

We denote the policy LAOC on the basis of the ML policy $\tilde{\pi}$ and the action set $\mathcal{U}_{\lambda, h}$ as

$$\pi_\lambda(s_h) = m(\tilde{\pi}(s_h), \mathcal{U}_{\lambda, h}), \quad (28)$$

where s_h is the ML input at round h , and m is the projection function in (8) or the linear function in (9). By directly applying the ML policy $\tilde{\pi}$ without projection or linear operations, we get the action sequence $\{u_h^\dagger, h \in [H]\}$ and the state sequence $\{\tilde{x}_h, h \in [H]\}$, and the corresponding ML inputs (which include \tilde{x}_h) are denoted as \tilde{s}_h .

Lemma D.1. *Given two constants $\lambda_1 > 0$ and $\lambda_0 \in (0, \lambda)$, if the potential function is designed as $\phi_h(u) = q_h \|f_h(x_h, u) - f_t(x_h^\dagger, u_h^\dagger)\|^2$ with $q_h \geq 0$ satisfying $2\sigma_x^2 q_h \leq q_{h-1} - (1 + \frac{1}{\lambda_0}) \frac{\beta}{2}$ for $h \in [H-1]$, $q_H = 0$, then u_h is in the competitive action set (7) if*

$$\|u_h - u_h^\dagger\|^2 \leq G r_h^\dagger,$$

where r_h^\dagger is the risk of the prior at time h , and $G = \frac{(\lambda - \lambda_0)}{(1 + \frac{1}{\lambda_0}) \frac{\beta}{2} + \frac{C_2}{C_2 - 1} q_h \sigma_u^2}$.

PROOF. Note that at time $h-1$, the competitiveness constraint holds as

$$R_{h-1} + \phi_{h-1}(u_{h-1}) \leq (1 + \lambda) \left(R_{h-1}^\dagger \right), \quad (29)$$

and the sufficient condition for $u_h \in \mathcal{U}_{\lambda, h}$ is

$$r_h(x_h, u_h) + \phi_h(u_h) - \phi_{h-1}(u_{h-1}) \leq (1 + \lambda) r_h(x_h^\dagger, u_h^\dagger). \quad (30)$$

Given $\lambda_0 \in (0, \lambda)$, subtracting $(1 + \lambda_0)r_h(x_h^\dagger, u_h^\dagger)$ by both sides and by Lemma C.1, we get the sufficient condition that (30) holds if

$$\left(\left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} - q_{h-1} \right) \|x_h - x_h^\dagger\|^2 + \left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} \|u_h - u_h^\dagger\|^2 + q_h \|x_{h+1} - x_{h+1}^\dagger\|^2 \leq (\lambda - \lambda_0)r_h^\dagger. \quad (31)$$

By the Lipschitz continuity of f and the smoothness of squared norm, we have by Lemma C.1

$$\begin{aligned} \|x_{h+1} - x_{h+1}^\dagger\|^2 &= \|f_h(x_h, u_h) - f(x_h^\dagger, u_h^\dagger)\|^2 \\ &\leq C_2 \sigma_x^2 \|x_h - x_h^\dagger\|^2 + \frac{C_2}{C_2 - 1} \sigma_u^2 \|u_h - u_h^\dagger\|^2. \end{aligned}$$

Thus, we further get the sufficient condition of (31) as

$$\begin{aligned} &\left(\left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} + C_2 q_h \sigma_x^2 - q_{h-1} \right) \|x_h - x_h^\dagger\|^2 + \\ &\left(\left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} + \frac{C_2}{C_2 - 1} q_h \sigma_u^2 \right) \|u_h - u_h^\dagger\|^2 \leq (\lambda - \lambda_0)r_h^\dagger. \end{aligned} \quad (32)$$

By the condition that $q_{h-1} \geq \left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} + C_2 q_h \sigma_x^2$, we get the following:

$$\left(\left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} + \frac{C_2}{C_2 - 1} q_h \sigma_u^2 \right) \|u_h - u_h^\dagger\|^2 \leq (\lambda - \lambda_0)r_h^\dagger, \quad (33)$$

which implies the sufficient condition in this lemma. \square

Lemma D.2. *With a reservation function satisfying the condition in Lemma D.1, $\xi_h(s_h) = \|\pi_\lambda(s_h) - \tilde{\pi}(s_h)\| = \|m(\tilde{\pi}(s_h), \mathcal{U}_{\lambda, h}) - \tilde{\pi}(s_h)\|$ with m being the projection function in (28) is bounded by*

$$\xi_h(s_h) \leq L_\pi \|x_h - \tilde{x}_h\| + \left[\delta_h - (\sqrt{1 + \lambda} - 1)^2 G r_h^\dagger \right]^+,$$

where $G = \frac{2}{L_c \left(1 + \frac{C_2}{C_2 - 1} \sigma_u^2 (1 - (C_2 \sigma_x^2)^{H-h}) / (1 - C_2 \sigma_x^2)\right)}$ and $\delta_h = \|\tilde{\pi}(\tilde{s}_h) - \pi^\dagger(s_h^\dagger)\|$.

PROOF. We choose $q_h = \left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} \sum_{i=0}^{H-h-1} (C_2 \sigma_x^2)^i$ given any $\lambda_0 \in (0, \lambda)$. The choice of q_h satisfies the requirement for q_h in Lemma D.1 and the sufficient condition becomes

$$\begin{aligned} \|u_h - u_h^\dagger\|^2 &\leq \frac{\lambda - \lambda_0}{\left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} + \frac{C_2}{C_2 - 1} \sigma_u^2 \sum_{i=0}^{H-h-1} (C_2 \sigma_x^2)^i} r_h^\dagger \\ &= \frac{\lambda - \lambda_0}{\left(1 + \frac{1}{\lambda_0}\right) \frac{\beta}{2} + \frac{C_2}{C_2 - 1} \sigma_u^2 (1 - (C_2 \sigma_x^2)^{H-h}) / (1 - C_2 \sigma_x^2)} r_h^\dagger. \end{aligned} \quad (34)$$

By optimally choosing $\lambda_0 = \sqrt{1 + \lambda} - 1$, we have

$$q_h = \left(1 + \frac{1}{\sqrt{1 + \lambda} - 1}\right) \frac{\beta}{2} \sum_{i=0}^{H-h-1} (C_2 \sigma_x^2)^i,$$

and

$$\|u_h - u_h^\dagger\|^2 \leq \frac{\frac{2}{\beta} (\sqrt{1 + \lambda} - 1)^2 r_h^\dagger}{1 + \frac{C_2}{C_2 - 1} \sigma_u^2 (1 - (C_2 \sigma_x^2)^{H-h}) / (1 - C_2 \sigma_x^2)}. \quad (35)$$

Since $\|u_h - u_h^\dagger\| \leq A$ where A is the size of the action set, we get the sufficient condition that an action belongs to safe action set (7) as

$$\begin{aligned} \|u_h - u_h^\dagger\| &\leq \frac{\frac{2}{L_c} (\sqrt{1 + \lambda} - 1)^2 r_h^\dagger}{1 + \frac{C_2}{C_2 - 1} \sigma_u^2 (1 - (C_2 \sigma_x^2)^{H-h}) / (1 - C_2 \sigma_x^2)} \\ &= (\sqrt{1 + \lambda} - 1)^2 G r_h^\dagger. \end{aligned} \quad (36)$$

Let $G' = (\sqrt{1 + \lambda} - 1)^2 G$. We denote the action projected from $\tilde{\pi}(s_h)$ to the norm ball $\mathcal{B}(u_h^\dagger, G' r_h^\dagger)$ as $\pi_\lambda^\dagger(s_h)$. We have $\pi_\lambda^\dagger(s_h) = \tilde{\pi}(s_h)$ if $\tilde{\pi}(s_h) \in \mathcal{B}(u_h^\dagger, G' r_h^\dagger)$. And if $\tilde{\pi}(s_h) \notin \mathcal{B}(u_h^\dagger, G' r_h^\dagger)$, we have $\pi_\lambda^\dagger(s_h) = u_h^\dagger + G' r_h^\dagger \frac{\tilde{\pi}(s_h) - u_h^\dagger}{\|\tilde{\pi}(s_h) - u_h^\dagger\|}$. Since $\mathcal{U}_{\lambda, h}$ is a close set, the norm ball $\mathcal{B}(u_h^\dagger, G' r_h^\dagger) \subset \mathcal{U}_{\lambda, h}$ thanks to Lemma D.1, we have

$$\begin{aligned} \xi_h(s_h) &= \|\pi_\lambda(s_h) - \tilde{\pi}(s_h)\| \leq \|\pi_\lambda^\dagger(s_h) - \tilde{\pi}(s_h)\| \\ &= \left[\|\tilde{\pi}(s_h) - u_h^\dagger\| - G' r_h^\dagger \right]^+ \\ &\leq \|\tilde{\pi}(\tilde{s}_h) - \tilde{\pi}(s_h)\| + \left[\|\tilde{\pi}(\tilde{s}_h) - u_h^\dagger\| - G' r_h^\dagger \right]^+ \\ &\leq L_\pi \|x_h - \tilde{x}_h\| + \left[\delta_h - G' r_h^\dagger \right]^+, \end{aligned} \quad (37)$$

where the first inequality is because π_λ applies the projection or linear operation m on the ML predictions, the second equality holds because if $\tilde{\pi}(s_h) \notin \mathcal{B}(u_h^\dagger, G' r_h^\dagger)$, $\|\pi_\lambda^\dagger(s_h) - \tilde{\pi}(s_h)\| = \|u_h^\dagger - \tilde{\pi}(s_h) - G' r_h^\dagger \frac{\tilde{\pi}(s_h) - u_h^\dagger}{\|\tilde{\pi}(s_h) - u_h^\dagger\|}\| = \|\tilde{\pi}(s_h) - u_h^\dagger\| - G' r_h^\dagger$ and if $\tilde{\pi}(s_h) \in \mathcal{B}(u_h^\dagger, G' r_h^\dagger)$, $\|\pi_\lambda^\dagger(s_h) - \tilde{\pi}(s_h)\| = 0$, the second inequality holds by the triangle inequality, and the last inequality holds by the Lipschitz continuity of the policy π^* and $\|\tilde{s}_h - s_h\| = \|\tilde{x}_h - x_h\|$ for the same context instance. \square

Lemma D.3. *With a reservation function satisfying the condition in Lemma D.1, the difference of the states with respect to the policy π_λ and the policy $\tilde{\pi}$ is bounded as*

$$\|\tilde{x}_h - x_h\| \leq \sum_{i=0}^{h-1} (\sigma_x + 2\sigma_u L_\pi)^{h-i-1} \sigma_u \left[\eta_i - (\sqrt{1 + \lambda} - 1)^2 G r_i^\dagger \right]^+,$$

where $G = \frac{2}{L_c \left(1 + \frac{C_2}{C_2 - 1} \sigma_u^2 (1 - (C_2 \sigma_x^2)^{H-h}) / (1 - C_2 \sigma_x^2)\right)}$.

PROOF. By the state dynamic function and Lipschitz continuity, we have

$$\begin{aligned} \|\tilde{x}_h - x_h\| &= \|f(\tilde{x}_{h-1}^\dagger, \tilde{\pi}(\tilde{s}_{h-1})) - f(x_{h-1}, \pi_\lambda(s_{h-1}))\| \\ &\leq \sigma_x \|\tilde{x}_{h-1}^\dagger - x_{h-1}\| + \sigma_u \|\tilde{\pi}(\tilde{s}_{h-1}) - \pi_\lambda(s_{h-1})\| \\ &\leq \sigma_x \|\tilde{x}_{h-1}^\dagger - x_{h-1}\| + \sigma_u \|\tilde{\pi}(\tilde{s}_{h-1}) - \tilde{\pi}(s_{h-1})\| \\ &\quad + \sigma_u \|\tilde{\pi}(s_{h-1}) - \pi_\lambda(s_{h-1})\| \\ &\leq (\sigma_x + \sigma_u L_\pi) \|\tilde{x}_{h-1}^\dagger - x_{h-1}\| + \sigma_u \xi_{h-1}(s_{h-1}), \end{aligned} \quad (38)$$

where the second inequality holds by the triangle inequality, and the last inequality holds by the Lipschitz continuity of the policy $\tilde{\pi}$ and $\|\tilde{s}_h - s_h\| = \|\tilde{x}_h - x_h\|$ for the same context instance.

Applying Lemma D.2 for $\xi_{h-1}(s_{h-1})$, we further have

$$\begin{aligned} \|\tilde{x}_h - x_h\| &\leq (\sigma_x + 2\sigma_u L_\pi) \|\tilde{x}_{h-1}^\dagger - x_{h-1}\| \\ &\quad + \sigma_u \left[\eta_{h-1} - (\sqrt{1 + \lambda} - 1)^2 G r_{h-1}^\dagger \right]^+. \end{aligned} \quad (39)$$

Iteratively applying (39), we have

$$\|\tilde{x}_h - x_h\| \leq \sum_{i=0}^{h-1} (\sigma_x + 2\sigma_u L_\pi)^{h-i-1} \sigma_u \left[\eta_i - (\sqrt{1+\lambda} - 1)^2 G r_i^\dagger \right]^+. \quad (40)$$

□

Proof of Theorem 4.4

PROOF. Now we are ready to bound the difference of expected costs of LAOC and the pure ML policy $\tilde{\pi}$ which is

$$\begin{aligned} & \mathbb{E}_{y_{0:H}} \left[J_H^{\pi_\lambda}(y_{0:H}) \right] - \mathbb{E}_{y_{0:H}} \left[J_H^{\tilde{\pi}}(y_{0:H}) \right] \\ &= \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^H c_h(x_h, m(\tilde{\pi}(s_h), \mathcal{U}_{\lambda,h})) - c_h(\tilde{x}_h, \tilde{\pi}(\tilde{s}_h)) \right]. \end{aligned} \quad (41)$$

We can bound this difference as

$$\begin{aligned} & \mathbb{E}_{y_{0:H}} \left[J_H^{\pi_\lambda}(y_{0:H}) \right] - \mathbb{E}_{y_{0:H}} \left[J_H^{\tilde{\pi}}(y_{0:H}) \right] \\ &= \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^H c_h(x_h, m(\tilde{\pi}(s_h), \mathcal{U}_{\lambda,h})) - c_h(x_h, \tilde{\pi}(s_h)) \right. \\ & \quad \left. + c_h(x_h, \tilde{\pi}(s_h)) - c_h(\tilde{x}_h, \tilde{\pi}(\tilde{s}_h)) \right] \\ &\leq L_c \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^H \|m(\tilde{\pi}(s_h), \mathcal{U}_{\lambda,h}) - \tilde{\pi}(\tilde{s}_h)\| + (1+L_\pi)\|x_h - \tilde{x}_h\| \right] \\ &\leq L_c \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^H \xi_h(s_h) + (1+2L_\pi)\|x_h - \tilde{x}_h\| \right], \end{aligned} \quad (42)$$

where the first inequality holds because the cost functions c_h are L_c -Lipschitz continuous, $\tilde{\pi}$ is L_π -Lipschitz and $\tilde{s}_h - s_h = \tilde{x}_h - x_h$ for the same context instance. and the second equality is due to the definition of $\xi_h(s_h) = \|m(\tilde{\pi}(s_h), \mathcal{U}_{\lambda,h}) - \tilde{\pi}(s_h)\|$ in Lemma D.2 and $\|\tilde{\pi}(\tilde{s}_h) - \tilde{\pi}(s_h)\| \leq L_\pi \|\tilde{s}_h - s_h\| = L_\pi \|x_h - \tilde{x}_h\|$.

By Lemma D.2, we can further bound the expected cost difference as

$$\begin{aligned} & \mathbb{E}_{y_{0:H}} \left[J_H^{\pi_\lambda}(y_{0:H}) \right] - \mathbb{E}_{y_{0:H}} \left[J_H^{\tilde{\pi}}(y_{0:H}) \right] \\ &\leq L_c \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^H \xi_h(s_h) + (1+2L_\pi)\|x_h - \tilde{x}_h\| \right] \\ &\leq L_c \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^{H-1} \left[\delta_h - G' r_h^\dagger \right]^+ + (1+2L_\pi) \sum_{h=0}^H \|x_h - \tilde{x}_h\| \right], \end{aligned} \quad (43)$$

where $G' = (\sqrt{1+\lambda} - 1)^2 G$, $\delta_h = \|\tilde{\pi}(\tilde{s}_h) - \pi^\dagger(s_h^\dagger)\|$, and $\xi_H(s_H) = 0$ as there is no action at round H .

By Lemma D.3, the expected cost is bounded as

$$\begin{aligned} & \mathbb{E}_{y_{0:H}} \left[J_H^{\pi_\lambda}(y_{0:H}) \right] - \mathbb{E}_{y_{0:H}} \left[J_H^{\tilde{\pi}}(y_{0:H}) \right] \\ &\leq L_c \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^{H-1} \left[\delta_h - G' r_h^\dagger \right]^+ + \right. \\ & \quad \left. (1+2L_\pi) \sum_{h=1}^H \sum_{i=0}^{h-1} (\sigma_x + 2\sigma_u L_\pi)^{h-i-1} \sigma_u \left[\eta_i - G' r_i^\dagger \right]^+ \right] \\ &\leq L_c \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^{H-1} \left[\delta_h - G' r_h^\dagger \right]^+ + \right. \\ & \quad \left. (1+2L_\pi) \sigma_u \sum_{h=0}^{H-1} \left[\delta_h - G' r_h^\dagger \right]^+ \sum_{i=h}^{H-1} (\sigma_x + 2\sigma_u L_\pi)^{h-i-1} \right] \\ &\leq B \mathbb{E}_{y_{0:H}} \left[\sum_{h=0}^{H-1} \left[\delta_h - (\sqrt{1+\lambda} - 1)^2 G r_h^\dagger \right]^+ \right] \end{aligned} \quad (44)$$

where $B = L_c \left(1 + (1+2L_\pi) \sigma_u \sum_{i=0}^{H-1} (\sigma_x + 2\sigma_u L_\pi)^{h-i-1} \right)$.

The reservation function in Lemma D.2 meet the requirements in Proposition 4.2 by choosing some proper constants ρ , C_1 and C_2 . □

E Proof of Theorem 4.5

PROOF. Since the policy $\pi_\lambda^{(n)}$ is one from the constrained policy set Π_λ , we apply the statistical generalization theorem in [15] and get with probability at least $1 - \delta$, $\delta \in (0, 1)$,

$$\left| \mathbb{E} \left[J_H^{\pi_\lambda^{(n)}} \right] - \frac{1}{n} \sum_{t=1}^n J_H^{\pi_\lambda^{(n)}}(y_{0:H}^{(t)}) \right| \leq 4HP \sqrt{\frac{2}{n} \ln \frac{4N(\epsilon, \Pi_\lambda, \hat{L}_1^n)}{\delta}}, \quad (45)$$

where $N(\epsilon, \Pi_\lambda, \hat{L}_1^n)$ is the ϵ -covering number of the competitive policy space Π_λ with L_1 -norm as the distance measure: the distance of two functions π and π' is $\|\pi - \pi'\|_{\hat{L}_1^n} = \frac{1}{n} \sum_{t=1}^n \|\pi(s^{(t)}) - \pi'(s^{(t)})\|_1$.

By Eqn. (10), we have $\frac{1}{n} \sum_{t=1}^n J_H^{\pi_\lambda^{(n)}}(y_{0:H}^{(t)}) \leq \frac{1}{n} \sum_{t=1}^n J_H^{\pi_\lambda^*}(y_{0:H}^{(t)})$. Thus, we have

$$\begin{aligned} \mathbb{E} \left[J_H^{\pi_\lambda^{(n)}} \right] &\leq \frac{1}{n} \sum_{t=1}^n J_H^{\pi_\lambda^*}(y_{0:H}^{(t)}) + 4HP \sqrt{\frac{2}{n} \ln \frac{4N(\epsilon, \Pi_\lambda, \hat{L}_1^n)}{\delta}} \\ &\leq \mathbb{E} \left[J_H^{\pi_\lambda^*} \right] + 8HP \sqrt{\frac{2}{n} \ln \frac{4N(\epsilon, \Pi_\lambda, \hat{L}_1^n)}{\delta}}, \end{aligned} \quad (46)$$

where the last inequality holds by applying the generalization theorem in [15]. By Eqn.(44), we have

$$\begin{aligned} \mathbb{E} \left[J_H^{\pi_\lambda^{(n)}} \right] &\leq \mathbb{E} \left[J_H^{\pi_\lambda^*} \right] + B \mathbb{E} \left[\sum_{h=0}^{H-1} \left[\delta_h - (\sqrt{1+\lambda} - 1)^2 G r_h^\dagger \right]^+ \right] \\ & \quad + O \left(\sqrt{\frac{1}{n} \ln \frac{N(\epsilon, \Pi_\lambda, \hat{L}_1^n)}{\delta}} \right), \end{aligned} \quad (47)$$

where O notation indicates the increasing with episode length H and maximum loss value P □