A Convex Optimization Approach to Distributionally Robust Markov Decision Processes With Wasserstein Distance

Insoon Yang, *Member, IEEE*

***Abstract*—We consider the problem of constructing control policies that are robust against distribution errors in the model parameters of Markov decision processes. The Wasserstein metric is used to model the *ambiguity set* of admissible distributions. We prove the existence and optimality of Markov policies and develop convex optimization-based tools to compute and analyze the policies. Our methods, which are based on the Kantorovich convex relaxation and duality principle, have the following advantages. First, the proposed dual formulation of an associated Bellman equation resolves the infinite dimensionality issue that is inherent in its original formulation when the nominal distribution has a finite support. Second, our duality analysis identifies the structure of a worst-case distribution and provides a simple decentralized method for its construction. Third, a sensitivity analysis tool is developed to quantify the effect of ambiguity set parameters on the performance of distributionally robust policies. The effectiveness of our proposed tools is demonstrated through a human-centered air conditioning problem.**

***Index Terms*—Optimal control, stochastic systems, Markov processes, probability distribution, optimization, robustness.**

I. Introduction

THE STOCHASTIC and dynamic environments of many practical sequential decision-making problems cannot be perfectly modeled, which is partially due to inaccurate distributional information regarding uncertainties (e.g., [1] and [2]). To obtain a control strategy that is robust against uncertainties in model parameters such as transition probabilities and rewards in Markov decision processes (MDP), *robust* MDP formulations have been proposed [3]–[6]. However, these methods do not incorporate *a priori* distributional information about uncertainties because model parameters must be contained in a known set of possible parameter realizations. Therefore, robust MDP approaches often produce conservative control policies [7]. To overcome this limitation, a *distributionally robust* MDP formulation has recently been proposed to maximize the worst-case expected reward, assuming that the distribution of uncertain parameters is not fully known but lies in a so-called *ambiguity set* of probability distributions [7], [8]. For continuous state space models, [9] proposes a semidefinite programming approach to computing optimal linear feedback strategies in a linear-quadratic setting. A distributionally robust safety specification tool is developed in [10] to handle a probabilistic safety constraint that allows for distributional errors of uncertain variables. All these methods adopt ambiguity sets of distributions with moment and/or confidence interval constraints.

We consider a distributionally robust MDP problem by employing a different ambiguity set modeling approach using the Wasserstein metric [11]. In single-stage distributionally robust optimization problems, this statistical distance approach has been shown to be particularly useful when the volume of data is too small to reliably estimate the moments of an underlying distribution [12]–[14]. The main contributions of this letter are as follows. First, we prove the optimality of Markov policies and their existence in distributionally robust MDP problems with Wasserstein distance. Second, we propose a convex formulation of an associated Bellman equation using the Kantorovich duality principle [15] and the strong duality result of Gao and Kleywegt [14]. In particular, we completely resolve the infinite dimensionality issue in the original Bellman equation without sacrificing optimality when the nominal distribution of the ambiguity set has a finite support. Our two different formulations of the Bellman equation can be efficiently solved by distributed and centralized convex optimization methods, respectively. Third, we identify the structure of a worst-case distribution. This structure allows us to design a simple decentralized method for constructing the worst-case distribution. Fourth, we develop a sensitivity analysis tool by combining the envelope theorem and Kantorovich duality. This tool is useful in quantifying the effect of the parameters in the Wasserstein ball-based ambiguity set on the maximal expected reward. The effectiveness of the proposed convex optimization-based tools is demonstrated through an example of controlling air conditioners under ambiguous user preferences and behaviors.

Manuscript received March 6, 2017; revised April 16, 2017; accepted May 20, 2017. Date of publication June 5, 2017; date of current version June 14, 2017. This work was supported by the NSF CRII: CPS under Grant CNS-1657100. Recommended by Senior Editor G. Yin.

The author is with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2563 USA (e-mail: insoonya@usc.edu).

Digital Object Identifier 10.1109/LCSYS.2017.2711553

The rest of this letter is organized as follows. In Section II, we introduce a dynamic game formulation of distributionally robust MDP problems with Wasserstein distance. Section III contains all the main results of this letter. In Section IV, we apply the proposed convex optimization-based tools to a human-centered air conditioning problem.

A. Preliminaries and Notation

A finite-horizon Markov decision process (MDP) is defined as a 5-tuple $\langle T, \mathbb{S}, \mathbb{A}, \mathbf{p}, \mathbf{r}\rangle$, where T is the time horizon, $\mathbb{S}$ is the set of states, and $\mathbb{A}_{\boldsymbol{s}}$ is the set of actions given the state $\boldsymbol{s} \in \mathbb{S}$. These two sets are assumed to be finite. The matrices $\mathbf{p} \in \mathbb{R}^{|\mathbb{S}||\mathbb{A}_{\boldsymbol{s}}|\times|\mathbb{S}|}$ and $\mathbf{r} \in \mathbb{R}^{|\mathbb{A}_{\boldsymbol{s}}|\times|\mathbb{S}|}$ contain information about the transition probability and the reward, respectively. We let $P(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$ be the transition probability that the next state is $\boldsymbol{s}'$ given the current state-action pair $(\boldsymbol{s}, \boldsymbol{a}) \in \mathbb{S} \times \mathbb{A}_{\boldsymbol{s}}$, and $r(\boldsymbol{s}, \boldsymbol{a})$ be the reward for the state-action pair $(\boldsymbol{s}, \boldsymbol{a}) \in \mathbb{S} \times \mathbb{A}_{\boldsymbol{s}}$. The reward values are assumed to be bounded. The vector $\mathbf{r}_{\boldsymbol{s}}$ denotes the column of $\mathbf{r}$ associated with the state $\boldsymbol{s}$, where $\mathbf{r}_{\boldsymbol{s}} \coloneqq \{r(\boldsymbol{s}, \boldsymbol{a})\}_{\boldsymbol{a}\in\mathbb{A}_{\boldsymbol{s}}} \in \mathbb{R}^{|\mathbb{A}_{\boldsymbol{s}}|}$. The vector $\mathbf{p}_{\boldsymbol{s}}$ denotes the columns of $\mathbf{p}$ associated with the state $\boldsymbol{s}$, where $\mathbf{p}_{\boldsymbol{s}} \coloneqq \{\{P(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})\}_{\boldsymbol{s}'\in\mathbb{S}}\}_{\boldsymbol{a}\in\mathbb{A}_{\boldsymbol{s}}} \in \mathbb{R}^{|\mathbb{S}||\mathbb{A}_{\boldsymbol{s}}|}$.[1] The state at stage t is denoted as $s_t \in \mathbb{S}$. Given a Borel space X, $\mathcal{P}(X)$ denotes the set of Borel probability measures on X. Finally, we let $\mathcal{T} \coloneqq \{1, \ldots, T-1\}$ to denote the set of stages up to $T-1$.

II. The Setup

A. Dynamic Game Formulation

We consider a special class of MDPs in which the transition probability $\mathbf{p}_{\boldsymbol{s}}$ and the reward $\mathbf{r}_{\boldsymbol{s}}$ are not completely known; however, their joint distribution is assumed to be contained in a so-called *ambiguity set*, $\mathbb{D}_{\boldsymbol{s}} \subseteq \mathcal{P}(\mathbb{R}^{|\mathbb{S}||\mathbb{A}_{\boldsymbol{s}}|+|\mathbb{A}_{\boldsymbol{s}}|})$, which is given. Our goal is to construct a control policy that maximizes the worst-case expected total reward under distributional constraints characterized by the ambiguity set $\mathbb{D}_{\boldsymbol{s}}$. We consider a dynamic game formulation in which Player I determines a control policy to maximize the reward while Player II selects the joint distribution μ_t of $(\mathbf{p}_{s_t}, \mathbf{r}_{s_t})$ to minimize the total expected reward.

Let $h_t \coloneqq (s_1, a_1, \mu_1, \ldots, s_{t-1}, a_{t-1}, \mu_{t-1}, s_t)$ be the *history* at stage t and H_t denote the set of all histories at stage t. The set of all history-dependent randomized control policies for Player I is denoted by Π. In other words, for a strategy $\pi \coloneqq (\pi_1, \ldots, \pi_{T-1}) \in \Pi$, we have that $\pi_t : H_t \to \Delta(\mathbb{A}_{s_t})$, where $\Delta(X)$ denotes the probability simplex on a set X. Now, let $h_t^e \coloneqq (s_1, a_1, \mu_1, \ldots, s_{t-1}, a_{t-1}, \mu_{t-1}, s_t, a_t)$ be the extended history at stage t and H_t^e denote the set of all extended histories at stage t. The set of Player II's admissible policies is defined as $\Gamma \coloneqq \{\gamma \coloneqq (\gamma_1, \ldots, \gamma_{T-1}) \mid \gamma_t : H_t^e \to \mathbb{D}_{s_t}\ \forall t \in \mathcal{T}\}$, which encodes the ambiguity set $\mathbb{D}_{s_t}$.

[1] Similarly to [3], [8], and [7], the MDP is assumed to be non-stationary in the sense that each time a state is visited, it can have a different realization of $(\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}})$. The proposed methods are also valid with time-varying rewards and transition probabilities. However, for notational simplicity, we suppress the time-dependency of $(\mathbf{p}_{\boldsymbol{s},t}, \mathbf{r}_{\boldsymbol{s},t})$.

Given a parameter pair $(\mathbf{p}, \mathbf{r})$ and a strategy pair $(\pi, \gamma) \in \Pi \times \Gamma$, we set the expected total reward as

$$R_{\boldsymbol{s}}[\pi, \gamma] \coloneqq \mathbb{E}^{\pi,\gamma}\left[\sum_{t=1}^{T-1} r(s_t, a_t) + q(s_T)\middle| s_1 = \boldsymbol{s}\right],$$

where $\mathbb{E}^{\pi,\gamma}$ denotes the expectation taken with respect to the probability measure induced by the strategy pair (π, γ), and $\boldsymbol{s}$ is the initial state, which is deterministic. Here, q is a terminal reward function, which is assumed to be bounded. Our desired control policy can be obtained by solving the following zero-sum two-player dynamic game problem[2]:

$$\sup_{\pi\in\Pi}\inf_{\gamma\in\Gamma} R_{\boldsymbol{s}}[\pi, \gamma]. \tag{1}$$

If it exists, an optimal solution to this problem maximizes the total expected reward under the worst-possible strategy for the joint distribution μ_t of $(\mathbf{p}_{s_t}, \mathbf{r}_{s_t})$ for all t. This minimax formulation provides a control strategy such that the closed-loop system is robust against distributional errors within the feasibility set characterized by the constraints in $\mathbb{D}_{\boldsymbol{s}}$.

B. Ambiguity Sets With the Wasserstein Metric

The distributionally robust MDP problem does not rely on the notion of a known true underlying distribution but instead requires an ambiguity set, $\mathbb{D}_{\boldsymbol{s}}$, of admissible distributions. To model ambiguity sets, several methods have been proposed for single-stage stochastic programming problems, along with useful duality results, such as moment-based methods [1], [17]–[20] and statistical distance-based approaches (ϕ-divergence: [21]–[23], Prokhorov metric: [24], Wasserstein metric: [12]–[14]). From these, we employ Wasserstein distance-based ambiguity sets in the distributionally robust MDP setting (1).

We fix $\boldsymbol{s} \in \mathbb{S}$ and use the notation $x \coloneqq (\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}}) \in \mathcal{X}_{\boldsymbol{s}} \subset \mathbb{R}^{|\mathbb{S}||\mathbb{A}_{\boldsymbol{s}}|+|\mathbb{A}_{\boldsymbol{s}}|}$. We assume that its probability measure $\boldsymbol{\mu}$ with non-empty support $\mathcal{X}_{\boldsymbol{s}}$ belongs to the following *Wasserstein ball* with radius $\theta > 0$ centered at a nominal probability measure $\nu_{\boldsymbol{s}} \in \mathcal{P}(\mathcal{X}_{\boldsymbol{s}})$:

$$\mathbb{D}_{\boldsymbol{s}} \coloneqq \{\boldsymbol{\mu} \in \mathcal{P}(\mathcal{X}_{\boldsymbol{s}}) \mid W_p(\boldsymbol{\mu}, \nu_{\boldsymbol{s}}) \le \theta\}, \tag{2}$$

where $W_p(\boldsymbol{\mu}, \nu_{\boldsymbol{s}}) \coloneqq \min_{\kappa\in\mathcal{P}(\mathcal{X}_{\boldsymbol{s}}\times\mathcal{X}_{\boldsymbol{s}})}\{[\int_{\mathcal{X}_{\boldsymbol{s}}\times\mathcal{X}_{\boldsymbol{s}}} d(x, y)^p \mathrm{d}\kappa(x, y)]^{\frac{1}{p}} | \Pi^1\kappa = \boldsymbol{\mu}, \Pi^2\kappa = \nu_{\boldsymbol{s}}\}$ is the Wasserstein distance of order p between $\boldsymbol{\mu}$ and $\nu_{\boldsymbol{s}}$ with a metric d and $p \ge 1$. Here, $\Pi^i\kappa$ denotes the ith marginal of κ for $i = 1, 2$. We can interpret the Wasserstein distance between two measures as the minimum cost of redistributing mass from one to another using non-uniform perturbations [11]. Recently, distributionally robust stochastic optimization with Wasserstein distance has been empirically shown to resolve issues with ϕ-divergence, which does not address how close two points in the support are to each other [14].

III. Kantorovich Duality-Based Convex Formulation of Dynamic Programming Solutions

To solve the distributionally robust MDP problem (1), we first introduce the value function $v_t(\boldsymbol{s}) \coloneqq \sup_{\pi\in\Pi}\inf_{\gamma\in\Gamma}$

[2] In this letter, we focus on a finite-horizon problem. However, our results can be extended to discounted infinite-horizon cases by showing that an associated dynamic programming operator is a contraction mapping [16].

$\mathbb{E}^{\pi,\gamma}[\sum_{\tau=t}^{T} r(s_\tau, a_\tau) + q(s_T) \mid s_t = \boldsymbol{s}]$, which represents the expected reward-to-go. Applying the dynamic programming principle, we can derive the following Bellman equation [16], [25], [26]: for stages $t \in \mathcal{T}$,

$$v_t(\boldsymbol{s}) = \sup_{\boldsymbol{\pi}\in\Delta(\mathbb{A}_{\boldsymbol{s}})} \inf_{\boldsymbol{\mu}\in\mathbb{D}_{\boldsymbol{s}}} \int_{\mathcal{X}_{\boldsymbol{s}}} \sum_{\boldsymbol{a}\in\mathbb{A}_{\boldsymbol{s}}} \boldsymbol{\pi}[\boldsymbol{a}]\Big(r(\boldsymbol{s},\boldsymbol{a}) + \sum_{\boldsymbol{s}'\in\mathbb{S}} P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) v_{t+1}(\boldsymbol{s}')\Big) \mathrm{d}\boldsymbol{\mu}(\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}}), \tag{3}$$

and for stage T, $v_T(\boldsymbol{s}) = q(\boldsymbol{s})$. Note that the inner problem is infinite-dimensional due to the Wasserstein metric-based ambiguity set $\mathbb{D}_{\boldsymbol{s}}$.

A. Optimality of Markov Policies

We now show that the distributionally robust MDP problem admits a Markov control policy, which is optimal.

Theorem 1: For each $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathbb{S}$, there exists a function $\pi_t^{opt} : \mathbb{S} \to \Delta(\mathbb{A}_{\boldsymbol{s}})$ such that

$$v_t(\boldsymbol{s}) = \inf_{\boldsymbol{\mu}\in\mathbb{D}_{\boldsymbol{s}}} \int_{\mathcal{X}_{\boldsymbol{s}}} \sum_{\boldsymbol{a}\in\mathbb{A}_{\boldsymbol{s}}} \pi_t^{opt}(\boldsymbol{s})[\boldsymbol{a}]\Big(r(\boldsymbol{s},\boldsymbol{a}) + \sum_{\boldsymbol{s}'\in\mathbb{S}} P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) v_{t+1}(\boldsymbol{s}')\Big) \mathrm{d}\boldsymbol{\mu}(\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}}).$$

Proof: Fix $\boldsymbol{s} \in \mathbb{S}$. Let $\mathcal{L}(\boldsymbol{\pi}) := \inf_{\boldsymbol{\mu}\in\mathbb{D}_{\boldsymbol{s}}} \int_{\mathcal{X}_{\boldsymbol{s}}} \sum_{\boldsymbol{a}\in\mathbb{A}_{\boldsymbol{s}}} \boldsymbol{\pi}[\boldsymbol{a}] (r(\boldsymbol{s},\boldsymbol{a}) + \sum_{\boldsymbol{s}'\in\mathbb{S}} P(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) v_{t+1}(\boldsymbol{s}')) \mathrm{d}\boldsymbol{\mu}(\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}})$, which is concave because it is the pointwise infimum of linear functions. Note that $\mathcal{L}$ is proper because r and q are bounded. In addition, $\mathcal{L}$ is an upper semi-continuous function and thus it is closed. The Bellman equation can then be rewritten as $v_t(\boldsymbol{s}) = \sup_{\boldsymbol{\pi}\in\Delta(\mathbb{A}_{\boldsymbol{s}})} \mathcal{L}(\boldsymbol{\pi})$. The optimization problem admits an optimal solution because it maximizes a closed proper concave function, $\mathcal{L}$, over the probability simplex $\Delta(\mathbb{A}_{\boldsymbol{s}})$, which is a closed convex set. Setting $\pi_t^{opt}(\boldsymbol{s})$ as such an optimal solution, we can construct the desired function. ∎

Theorem 1 allows us to obtain Player I's optimal control policy as $\pi^{opt} := (\pi_1^{opt}, \ldots, \pi_{T-1}^{opt})$, which is Markov. However, the construction of this policy is computationally challenging because the Bellman equation involves an infinite-dimensional minimax problem for each $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathbb{S}$.

B. Convex Formulation Using Kantorovich Duality

To develop tractable convex optimization-based methods for solving the Bellman equation, we consider a dual formulation of the inner problem. Let $\mathbf{v}_{t+1} := \{v_{t+1}(\boldsymbol{s})\}_{\boldsymbol{s}\in\mathbb{S}} \in \mathbb{R}^{|\mathbb{S}|}$ and $\mathbf{V}_{t+1,\boldsymbol{s}} := [\mathbf{e}_1 \mathbf{v}_{t+1}^\top \cdots \mathbf{e}_{|\mathbb{A}_{\boldsymbol{s}}|} \mathbf{v}_{t+1}^\top] \in \mathbb{R}^{|\mathbb{A}_{\boldsymbol{s}}|\times|\mathbb{S}||\mathbb{A}_{\boldsymbol{s}}|}$, where $\mathbf{e}_i$ denotes the $|\mathbb{A}_{\boldsymbol{s}}|$-dimensional unit vector of which the ith entry is equal to 1. Based on Kantorovich duality [15], we can reformulate the Bellman equation as follows.

Theorem 2: The Bellman equation (3) associated with the distributionally robust MDP problem (1) is equivalent to

$$v_t(\boldsymbol{s}) = \max_{\boldsymbol{\pi}\in\Delta(\mathbb{A}_{\boldsymbol{s}}),\lambda\geq 0} f_{\boldsymbol{s}}(\boldsymbol{\pi}, \lambda; v_{t+1}), \tag{4}$$

for $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathbb{S}$ with the terminal condition $v_T(\boldsymbol{s}) = q(\boldsymbol{s})$, where $f_{\boldsymbol{s}}(\cdot, \cdot; v_{t+1}) : \Delta(\mathbb{A}_{\boldsymbol{s}}) \times \mathbb{R}_+ \to \mathbb{R}$ is defined as

$$f_{\boldsymbol{s}}(\boldsymbol{\pi}, \lambda; v_{t+1}) := -\lambda\theta^p + \int_{\mathcal{X}_{\boldsymbol{s}}} \inf_{\substack{x=(\mathbf{p}_{\boldsymbol{s}},\mathbf{r}_{\boldsymbol{s}})\\ \in\mathcal{X}_{\boldsymbol{s}}}} [\lambda d(x,y)^p + (\mathbf{r}_{\boldsymbol{s}} + \mathbf{V}_{t+1,\boldsymbol{s}}\mathbf{p}_{\boldsymbol{s}})^\top \boldsymbol{\pi}] \mathrm{d}\nu_{\boldsymbol{s}}(y).$$

In addition, $f_{\boldsymbol{s}}$ is jointly concave with respect to $(\boldsymbol{\pi}, \lambda)$.

Proof: The Kantorovich duality principle suggests that

$$W_p(\boldsymbol{\mu}, \nu_{\boldsymbol{s}})^p = \sup_{\varphi,\psi\in\Phi_d} \Big\{ \int_{\mathcal{X}_{\boldsymbol{s}}} \varphi(x) \mathrm{d}\boldsymbol{\mu}(x) + \int_{\mathcal{X}_{\boldsymbol{s}}} \psi(y) \mathrm{d}\nu_{\boldsymbol{s}}(y) \Big\},$$

where

$$\Phi_d := \{(\varphi, \psi) \in L^1(\mathrm{d}\boldsymbol{\mu}) \times L^1(\mathrm{d}\nu_{\boldsymbol{s}}) \mid \varphi(x) + \psi(y) \leq d(x,y)^p \ \forall x, y \in \mathcal{X}_{\boldsymbol{s}}\}. \tag{5}$$

Thus, for any $(\varphi, \psi) \in \Phi_d$, we have that $\psi(y) \leq \inf_{x\in\mathcal{X}_{\boldsymbol{s}}} d(x,y)^p - \varphi(x)$ for each $y \in \mathcal{X}_{\boldsymbol{s}}$. The Wasserstein ball (2) can then be expressed as

$$\mathbb{D}_{\boldsymbol{s}} = \Big\{ \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X}_{\boldsymbol{s}}) \Big| \int_{\mathcal{X}_{\boldsymbol{s}}} \varphi(x) \mathrm{d}\boldsymbol{\mu}(x) + \int_{\mathcal{X}_{\boldsymbol{s}}} \inf_{x\in\mathcal{X}_{\boldsymbol{s}}} [d(x,y)^p - \varphi(x)] \mathrm{d}\nu_{\boldsymbol{s}}(y) \leq \theta^p \ \forall \varphi \in L^1(\mathrm{d}\boldsymbol{\mu}) \Big\}.$$

Fix $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathbb{S}$ and recall that $x := (\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}})$. The following inequality holds for all $\varphi \in L^1(\mathrm{d}\boldsymbol{\mu})$:

$$\begin{aligned}
&\inf_{\boldsymbol{\mu}\in\mathbb{D}_{\boldsymbol{s}}} \int_{\mathcal{X}_{\boldsymbol{s}}} [(\mathbf{r}_{\boldsymbol{s}} + \mathbf{V}_{t+1,\boldsymbol{s}}\mathbf{p}_{\boldsymbol{s}})^\top \boldsymbol{\pi}] \mathrm{d}\boldsymbol{\mu}(\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}}) \\
&\geq \sup_{\lambda\geq 0} \inf_{\boldsymbol{\mu}\in\mathcal{P}(\mathcal{X}_{\boldsymbol{s}})} \Big\{ \int_{\mathcal{X}_{\boldsymbol{s}}} [(\mathbf{r}_{\boldsymbol{s}} + \mathbf{V}_{t+1,\boldsymbol{s}}\mathbf{p}_{\boldsymbol{s}})^\top \boldsymbol{\pi} + \lambda\varphi(x)] \mathrm{d}\boldsymbol{\mu}(x) \\
&\qquad + \int_{\mathcal{X}_{\boldsymbol{s}}} \inf_{x\in\mathcal{X}_{\boldsymbol{s}}} [\lambda d(x,y)^p - \lambda\varphi(x)] \mathrm{d}\nu_{\boldsymbol{s}}(y) - \lambda\theta^p \Big\},
\end{aligned}$$

where the left hand-side is a compact representation of the inner problem in the Bellman equation (3). Select $\varphi \in L^1(\mathrm{d}\boldsymbol{\mu})$ such that $\lambda\varphi = -(\mathbf{r}_{\boldsymbol{s}} + \mathbf{V}_{t+1,\boldsymbol{s}}\mathbf{p}_{\boldsymbol{s}})^\top \boldsymbol{\pi}$. Then, we have the following weak duality for the inner problem: $\inf_{\boldsymbol{\mu}\in\mathbb{D}_{\boldsymbol{s}}} \int_{\mathcal{X}_{\boldsymbol{s}}} [(\mathbf{r}_{\boldsymbol{s}} + \mathbf{V}_{t+1,\boldsymbol{s}}\mathbf{p}_{\boldsymbol{s}})^\top \boldsymbol{\pi}] \mathrm{d}\boldsymbol{\mu}(\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}}) \geq \sup_{\lambda\geq 0} f_{\boldsymbol{s}}(\boldsymbol{\pi}, \lambda)$, where we suppress the dependence of $f_{\boldsymbol{s}}$ on v_{t+1}. Using Theorem 1 of Gao and Kleywegt [14], we can further show that this inner problem is equivalent to $\sup_{\lambda\geq 0} f_{\boldsymbol{s}}(\boldsymbol{\pi}, \lambda)$ and that there exists a maximizer λ^{opt} that achieves the supremum. Therefore, the Bellman equation (3) can be rewritten as $v_t(\boldsymbol{s}) = \max_{\boldsymbol{\pi}\in\Delta(\mathbb{A}_{\boldsymbol{s}}),\lambda\geq 0} f_{\boldsymbol{s}}(\boldsymbol{\pi}, \lambda)$ since the outer maximization problem of the Bellman equation also admits an optimal solution by Theorem 1.

To show that $f_{\boldsymbol{s}}$ is concave, fix $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \Delta(\mathbb{A}_{\boldsymbol{s}})$, $\lambda_1, \lambda_2 \geq 0$, and $\rho \in (0, 1)$. Let $\boldsymbol{\pi}_\rho := \rho\boldsymbol{\pi}_1 + (1-\rho)\boldsymbol{\pi}_2$, and $\lambda_\rho := \rho\lambda_1 + (1-\rho)\lambda_2$. We then have $\boldsymbol{\pi}_\rho \in \Delta(\mathbb{A}_{\boldsymbol{s}})$ and $\lambda_\rho \geq 0$. For any $\epsilon > 0$, there exists $x_\rho \in \mathcal{X}_{\boldsymbol{s}}$ such that $f_{\boldsymbol{s}}(\boldsymbol{\pi}_\rho, \lambda_\rho) + \epsilon > -\lambda_\rho\theta^p + \int_{\mathcal{X}_{\boldsymbol{s}}} [\lambda_\rho d(x_\rho, y)^p + (\mathbf{r}_{\boldsymbol{s}} + \mathbf{V}_{t+1,\boldsymbol{s}}\mathbf{p}_{\boldsymbol{s}})^\top \boldsymbol{\pi}_\rho] \mathrm{d}\nu_{\boldsymbol{s}}(y)$. On the other hand, for $i = 1, 2$, we have $f_{\boldsymbol{s}}(\boldsymbol{\pi}_i, \lambda_i) \leq -\lambda_i\theta^p + \int_{\mathcal{X}_{\boldsymbol{s}}} [\lambda_i d(x_\rho, y)^p + (\mathbf{r}_{\boldsymbol{s}} + \mathbf{V}_{t+1,\boldsymbol{s}}\mathbf{p}_{\boldsymbol{s}})^\top \boldsymbol{\pi}_i] \mathrm{d}\nu_{\boldsymbol{s}}(y)$ since $x_\rho \in \mathcal{X}_{\boldsymbol{s}}$. Combining these inequalities, we obtain $f_{\boldsymbol{s}}(\boldsymbol{\pi}_\rho, \lambda_\rho) + \epsilon > \rho f_{\boldsymbol{s}}(\boldsymbol{\pi}_1, \lambda_1) + (1-\rho) f_{\boldsymbol{s}}(\boldsymbol{\pi}_2, \lambda_2)$. Letting ϵ tend to zero, we conclude that $f_{\boldsymbol{s}}$ is concave. ∎

Using the dual formulation of the Bellman equation in Theorem 2, we can solve the distributionally robust MDP problem via finite-dimensional convex programming. Note that the original infinite dimensionality issue is transferred to the

evaluation of $f_{\boldsymbol{s}}$, which requires us to solve a minimization problem for each $y \in \mathcal{X}_{\boldsymbol{s}}$. However, in many practical cases, particularly where there is a data-driven construction of the nominal distribution $\nu_{\boldsymbol{s}}$, it has a finite support. In such cases, we can completely remove the infinite dimensionality issue, as proposed in the following subsection.

C. Nominal Distribution With a Finite Support

Suppose that the nominal distribution $\nu_{\boldsymbol{s}}$ has a finite support, $\{\hat{x}_{\boldsymbol{s},1}, \ldots, \hat{x}_{\boldsymbol{s},N}\}$, $\hat{x}_{\boldsymbol{s},i} := (\hat{\mathbf{p}}_{\boldsymbol{s},i}, \hat{\mathbf{r}}_{\boldsymbol{s},i}) \in \mathbb{R}^{|\mathbb{S}||\mathbb{A}_{\boldsymbol{s}}|+|\mathbb{A}_{\boldsymbol{s}}|}$, i.e.,

$$\nu_{\boldsymbol{s}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{x}_{\boldsymbol{s},i}}, \tag{6}$$

where the indicator $\delta_{\hat{x}_{\boldsymbol{s},i}}(x)$ is equal to 1 if $x = \hat{x}_{\boldsymbol{s},i}$ and zero otherwise. Such a choice is useful in practice: for example, when choosing the nominal distribution $\nu_{\boldsymbol{s}}$ as an empirical distribution that is constructed from a finite number of data points or samples, the support of $\nu_{\boldsymbol{s}}$ is finite. In this case, $\hat{x}_{\boldsymbol{s},i}$'s can be selected as (a subset of) the data points.

If the nominal distribution is finitely supported, we can resolve the infinite dimensionality issue inherent in the inner minimization problem of the original Bellman equation (3).

Corollary 1 (Dual Bellman Equation I): Suppose that the nominal distribution $\nu_{\boldsymbol{s}}$ is given by (6) for each $\boldsymbol{s} \in \mathbb{S}$. Then, the Bellman equation (3) associated with the distributionally robust MDP problem (1) is equivalent to

$$v_t(\boldsymbol{s}) = \max_{\boldsymbol{\pi}\in\Delta(\mathbb{A}_{\boldsymbol{s}}),\lambda\geq 0} \frac{1}{N}\sum_{i=1}^{N} \hat{f}_{\boldsymbol{s},i}(\boldsymbol{\pi}, \lambda; v_{t+1}), \tag{7}$$

for $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathbb{S}$ with $v_T(\boldsymbol{s}) = q(\boldsymbol{s})$, where for $i = 1, \ldots, N$, $\hat{f}_{\boldsymbol{s},i}(\cdot, \cdot; v_{t+1}) : \Delta(\mathbb{A}_{\boldsymbol{s}}) \times \mathbb{R}_+ \to \mathbb{R}$ is defined as

$$\hat{f}_{\boldsymbol{s},i}(\boldsymbol{\pi}, \lambda; v_{t+1}) := -\lambda\theta^p + \inf_{x=(\mathbf{p}_{\boldsymbol{s}},\mathbf{r}_{\boldsymbol{s}})\in\mathcal{X}_{\boldsymbol{s}}} [\lambda d(x, \hat{x}_{\boldsymbol{s},i})^p + (\mathbf{r}_{\boldsymbol{s}} + \mathbf{V}_{t+1,\boldsymbol{s}}\mathbf{p}_{\boldsymbol{s}})^\top \boldsymbol{\pi}]. \tag{8}$$

In addition, $\hat{f}_{\boldsymbol{s},i}$ is jointly concave with respect to $(\boldsymbol{\pi}, \lambda)$.

By substituting $\nu_{\boldsymbol{s}}$ in Theorem 2 with (6), we can confirm that the statements in Corollary 1 are valid. Note that evaluating the objective function of (7) requires us to solve N finite-dimensional convex optimization problems, each of which is given in (8). Accordingly, the infinite-dimensionality issue in the original formulation is eliminated in this reformulated Bellman equation. Due to the additive structure of the objective function and the dependency of $\hat{f}_{\boldsymbol{s},i}$ on locally available data $\hat{x}_{\boldsymbol{s},i}$, it is also suitable to use distributed optimization methods to solve (7). However, for a centralized approach, the following equivalent formulation is useful:

Corollary 2 (Dual Bellman Equation II): Suppose that the nominal distribution $\nu_{\boldsymbol{s}}$ is given by (6). Then, the Bellman equation (3) associated with the distributionally robust MDP problem (1) is equivalent to

$$v_t(\boldsymbol{s}) = \max_{\boldsymbol{\pi}\in\Delta(\mathbb{A}_{\boldsymbol{s}})} \inf_{x\in\mathbb{B}_{\boldsymbol{s}}} \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\pi}^\top Q(v_{t+1})x_i \tag{9}$$

for $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathbb{S}$ with $v_T(\boldsymbol{s}) = q(\boldsymbol{s})$, where

$$\mathbb{B}_{\boldsymbol{s}} := \left\{(x_1, \ldots, x_N) \in \mathcal{X}_{\boldsymbol{s}}^N \mid \frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{x}_{\boldsymbol{s},i})^p \leq \theta^p\right\},$$

$$Q(v_{t+1}) := \begin{bmatrix}\mathbf{V}_{t+1} & I\end{bmatrix} \in \mathbb{R}^{|\mathbb{A}_{\boldsymbol{s}}|\times(|\mathbb{S}||\mathbb{A}_{\boldsymbol{s}}|+|\mathbb{A}_{\boldsymbol{s}}|)}.$$

Proof: We first claim that the right-hand side of (9) is less than or equal to $v_t(\boldsymbol{s})$. Due to the dual Bellman equation (7), for any $\epsilon > 0$, there exist $x_i^\epsilon := (\mathbf{p}_{\boldsymbol{s},i}^\epsilon, \mathbf{r}_{\boldsymbol{s},i}^\epsilon) \in \mathcal{X}_{\boldsymbol{s}}$ for $i = 1, \ldots, N$ such that

$$v_t(\boldsymbol{s}) + \epsilon > \max_{\boldsymbol{\pi}\in\Delta(\mathbb{A}_{\boldsymbol{s}}),\lambda\geq 0} \left\{\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\pi}^\top Q(v_{t+1})x_i^\epsilon + \lambda\Big(-\theta^p + \frac{1}{N}\sum_{i=1}^{N} d(x_i^\epsilon, \hat{x}_{\boldsymbol{s},i})^p\Big)\right\}.$$

If $\frac{1}{N}\sum_{i=1}^{N} d(x_i^\epsilon, \hat{x}_{\boldsymbol{s},i})^p > \theta^p$, then λ will be chosen to be $+\infty$ and this choice will result that $v_t(\boldsymbol{s}) = +\infty$. Thus, $\frac{1}{N}\sum_{i=1}^{N} d(x_i^\epsilon, \hat{x}_{\boldsymbol{s},i})^p \leq \theta^p$. Using the inequality above with this constraint, we have that

$$v_t(\boldsymbol{s}) + \epsilon > \max_{\boldsymbol{\pi}\in\Delta(\mathbb{A}_{\boldsymbol{s}}),\lambda\geq 0} \inf_{x\in\mathbb{B}_{\boldsymbol{s}}} \left\{\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\pi}^\top Q(v_{t+1})x_i + \lambda\Big(-\theta^p + \frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{x}_{\boldsymbol{s},i})^p\Big)\right\}.$$

We now notice that $\lambda \geq 0$ can be chosen to maximize $\lambda(-\theta^p + \frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{x}_{\boldsymbol{s},i})^p)$ and its maximum value is 0 for any $\{x_1, \ldots, x_N\}$ such that $\frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{x}_{\boldsymbol{s},i})^p \leq \theta^p$. Thus, the right-hand side of the equality above is less than or equal to the right-hand side of (9). Letting ϵ tend to zero, we obtain have that the right-hand side of (9) is less than equal to $v_t(\boldsymbol{s})$.

It now suffices to show that the right-hand side of (9) is bounded below by $v_t(\boldsymbol{s})$. We use weak duality to have that $\inf_{x\in\mathbb{B}_{\boldsymbol{s}}} \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\pi}^\top Q(v_{t+1})x_i \geq \sup_{\lambda\geq 0} \inf_{x\in\mathcal{X}_{\boldsymbol{s}}^N} \{\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\pi}^\top Q(v_{t+1})x_i + \lambda(\frac{1}{N}\sum_{i=1}^{N} d(x_i, \hat{x}_{\boldsymbol{s},i})^p - \theta^p)\}$. Note that the right-hand side of this inequality is equal to $\sup_{\lambda\geq 0} \frac{1}{N}\sum_{i=1}^{N} \hat{f}_{\boldsymbol{s},i}(\boldsymbol{\pi}, \lambda; v_{t+1})$. Using the dual Bellman equation (7), we conclude that the right-hand side of (9) is greater than equal to $v_t(\boldsymbol{s})$. ■

D. Decentralized Construction of a Worst-Case Distribution

We now consider the problem of constructing a worst-case probability distribution of $(\mathbf{p}_{\boldsymbol{s}}, \mathbf{r}_{\boldsymbol{s}})$, assuming that the nominal distribution has a finite support. The following proposition provides a simple structural characterization of the worst-case distribution.

Proposition 1: Suppose that the nominal distribution $\nu_{\boldsymbol{s}}$ is given by (6). If the inner minimization problem of (9) admits an optimal solution $(x_1^{opt}, \ldots, x_N^{opt})$, then

$$\boldsymbol{\mu}^{opt} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i^{opt}} \tag{10}$$

is a worst-case distribution, where $\boldsymbol{\pi}^{opt}$ is an optimal solution to the outer maximization problem of (9).

Proof: The original Bellman equation implies that

$$v_t(\boldsymbol{s}) \leq \int_{\mathcal{X}_{\boldsymbol{s}}} [(\boldsymbol{\pi}^{opt})^\top Q(v_{t+1})x]\mathrm{d}\boldsymbol{\mu}(x) \tag{11}$$

for any $\boldsymbol{\mu} \in \mathbb{D}_s$, where Q is given in Corollary 2. Using the Kantorovich duality principle, we have

$$W_p(\boldsymbol{\mu}^{opt}, \nu_s)^p = \sup_{\varphi,\psi \in \Phi_d} \frac{1}{N}\sum_{i=1}^{N} [\varphi(x_i^{opt}) + \psi(\hat{x}_{s,i})],$$

where the feasibility set Φ_d is given as (5). Since $\varphi(x) + \psi(y) \le d(x,y)^p$ for all $x, y \in \mathcal{X}_s$,

$$W_p(\boldsymbol{\mu}^{opt}, \nu_s)^p \le \frac{1}{N}\sum_{i=1}^{N} d(x_i^{opt}, \hat{x}_{s,i})^p \le \theta^p,$$

where the second inequality holds because of the constraint in the dual Bellman equation (9). This implies that $\boldsymbol{\mu}^{opt} \in \mathbb{D}_s$. Thus, the inequality (11) suggests that

$$\begin{aligned} v_t(\boldsymbol{s}) &\le \int_{\mathcal{X}_s} [(\boldsymbol{\pi}^{opt})^\top Q(v_{t+1})x] \mathrm{d}\boldsymbol{\mu}^{opt}(x) \\ &= \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{\pi}^{opt})^\top Q(v_{t+1}) x_i^{opt}. \end{aligned}$$

Due to Corollary 2, the right-hand side is equal to $v_t(\boldsymbol{s})$. Therefore, $\boldsymbol{\mu}^{opt}$ is a worst-case distribution. ■

This proposition is consistent with the fact that any worst-case distribution can have at most $2N$ support elements [14]. By combining Proposition 1, Corollary 1, and Corollary 2, we design an efficient method to construct the worst-case joint distribution of $(\mathbf{p}_s, \mathbf{r}_s)$. Given $(\boldsymbol{\pi}^{opt}, \lambda^{opt})$, the support element x_i^{opt} can be computed by solving the convex optimization problem (8) only with locally available data $\hat{x}_{s,i}$ due to Corollary 2. Note that this procedure is completely decentralized and parallelizable. Finally, $\boldsymbol{\mu}^{opt}$ can be constructed using Proposition 1.

E. Sensitivity Analysis via the Envelope Theorem

The radius θ of the Wasserstein ball (2) critically affects the effectiveness of the proposed distributionally robust MDP: when θ is too small, it would provide a control policy that is not sufficiently robust, while too large a θ would render an optimal strategy overly conservative. When selecting θ, there is a need for a sensitivity tool to display local behaviors of the value function with respect to θ without computing the value function for too many θ's. The precise effect of the order p (of Wasserstein distance) on the value function is also obscure. Using the envelope theorem [27], we show that both sensitivity values can be obtained from the solution result of the dual Bellman equation (4).

Proposition 2: For each $(t, \boldsymbol{s}) \in \mathcal{T} \times \mathbb{S}$, the sensitivity of the value function with respect to the radius θ of the Wasserstein ball (2) can be obtained as

$$\frac{\partial v_t(\boldsymbol{s})}{\partial \theta} = -(p\theta^{p-1})\lambda_{t,\boldsymbol{s}}^{opt} \le 0,$$

where $\lambda_{t,\boldsymbol{s}}^{opt} := \arg\max_{\lambda \ge 0} [\max_{\boldsymbol{\pi} \in \Delta(\mathbb{A}_s)} f_{\boldsymbol{s}}(\boldsymbol{\pi}, \lambda; v_{t+1})]$. In addition, the sensitivity of the value function with respect to the order p of Wasserstein distance can be computed as

$$\frac{\partial v_t(\boldsymbol{s})}{\partial p} = -(\theta^p \log p)\lambda_{t,\boldsymbol{s}}^{opt} \le 0.$$

Proof: Let $(\boldsymbol{\pi}_{t,\boldsymbol{s}}^{opt}, \lambda_{t,\boldsymbol{s}}^{opt})$ be an optimal solution of (4). Using the envelope theorem [27], we have that $\frac{\partial v_t(\boldsymbol{s})}{\partial \theta} = \frac{\partial f_{\boldsymbol{s}}(\boldsymbol{\pi}_{t,\boldsymbol{s}}^{opt}, \lambda_{t,\boldsymbol{s}}^{opt})}{\partial \theta} = -(p\theta^{p-1})\lambda_{t,\boldsymbol{s}}^{opt}$ because $p \ge 1$. This partial derivative is non-positive because $\lambda_{t,\boldsymbol{s}}^{opt} \ge 0$ and $\theta > 0$. Similarly, we can show that $\frac{\partial v_t(\boldsymbol{s})}{\partial p} = \frac{\partial f_{\boldsymbol{s}}(\boldsymbol{\pi}_{t,\boldsymbol{s}}^{opt}, \lambda_{t,\boldsymbol{s}}^{opt})}{\partial p} = -(\theta^p \log p)\lambda_{t,\boldsymbol{s}}^{opt} \le 0$, where the inequality holds since $p \ge 1$, $\lambda_{t,\boldsymbol{s}}^{opt} \ge 0$ and $\theta > 0$. ■

As the volume of the Wasserstein ball (2) increases with the radius θ and the order p, the value function decreases with respect to these two parameters. This sensitivity analysis is useful for examining local behaviors of the value function with respect to the radius and the order of Wasserstein distance, which are two important parameters in modeling the ambiguity set. Note that the proposed sensitivity tool is another useful byproduct of the Kantorovich duality-based convex formulation (4).

IV. Air Conditioning Under Ambiguous User Preferences and Behavioral Effects

Indoor temperatures of homes and buildings depend on several uncertainties including occupant behaviors, solar forcing, and outdoor temperatures. Furthermore, occupants often have uncertain preferences for comfortable indoor temperatures. A successful control method for air conditioning must effectively consider these uncertainties. However, it is difficult to obtain an accurate distribution of such uncertainties in practice. Thus, we use the proposed method to design a data-driven controller that is robust against errors in the joint distribution of (*i*) the reward vector $\mathbf{r}_s$ that models uncertainties in user preferences and (*ii*) the transition probability vector $\mathbf{p}_s$ in which users' behavioral uncertainties are considered.

We construct an MDP by simulating the following model of thermostatically controlled loads:

$$x_{t+1} = \kappa x_t + (1-\kappa)(\Theta - \eta RP u_t) + w_t,$$

where x_t, u_t and w_t represent the indoor temperature, control input and disturbance value at stage t.[3] We generate the samples of the reward vector by setting $r_t(\boldsymbol{s}, \boldsymbol{a}) := 0.95^t \times (r_1(\boldsymbol{s}) + \zeta r_2(\boldsymbol{a}))$, where $r_1(\boldsymbol{s}) = w - e^{(\boldsymbol{s}-\boldsymbol{s}^*)}\mathbf{1}_{\{\boldsymbol{s} \ge \boldsymbol{s}^*\}} - e^{(\boldsymbol{s}^*-\boldsymbol{s})}\mathbf{1}_{\{\boldsymbol{s} \le \boldsymbol{s}^*\}}$ represents the user satisfaction score regarding the indoor temperature $\boldsymbol{s}$ given the most preferred temperature $\boldsymbol{s}^* = 20.5$ (°C). The random variable w is normally distributed with a mean of 2 and a standard deviation of 0.2. On the other hand, $-r_2(\boldsymbol{a}) = c\boldsymbol{a}$ models the energy cost of air conditioning, where $c = \$0.01$ is the electricity price of unit kWh. The weight ζ is normally distributed with a mean of 2000 and a standard deviation of 200. The samples of the transition probability vectors are constructed by adding a normally distributed random variable with a mean of 0.05 and a standard deviation of 0.01 to the largest element of each column of the original transition probability matrix. The rest of the elements are scaled proportionally so that the sum of elements in each column remains 1. We choose these samples as the support elements $(\hat{\mathbf{p}}_{s,i}, \hat{\mathbf{r}}_{s,i})$ of the nominal distribution ν_s. We also let $d((\mathbf{p}_s, \mathbf{r}_s), (\mathbf{p}_s', \mathbf{r}_s'))^p := \|\mathbf{p}_s - \mathbf{p}_s'\|^2 + 10^{-3}\|\mathbf{r}_s - \mathbf{r}_s'\|^2$.

[3] A detailed explanation of the model and the choice of parameters in our simulations can be found in our previous work [10]. We choose the set of states as $\mathbb{S} := \{18 + 0.25 \times (i-1) \mid i = 1, \ldots, 21\}$ (unit: °C), the set of actions as $\mathbb{A}_s := \{1(\text{ON}), 0(\text{OFF})\}$ for all $\boldsymbol{s} \in \mathbb{S}$, and $\mathcal{T} := \{1, \ldots, 12\}$ with a 5-minute interval between two consecutive stages.

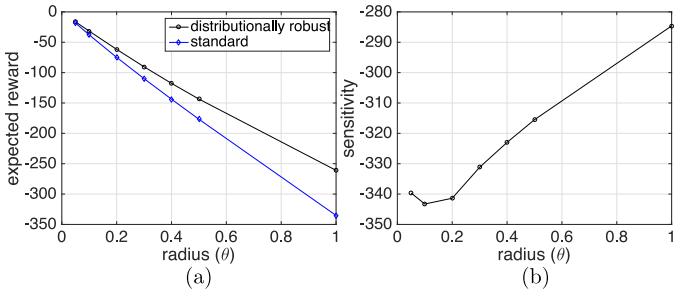

Fig. 1. (a) The worst-case expected rewards, and (b) the sensitivity of the worst-case reward with respect to θ when a distributionally robust strategy is employed.

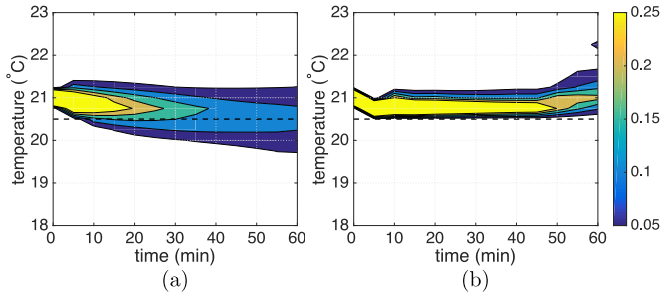

Fig. 2. The indoor temperature distribution controlled by (a) the proposed method and (b) non-robust method (e.g., [26]).

We illustrate our numerical experiment results by setting the initial condition at $s_1 = 21$ (°C). As shown in Fig. 1 (a), the worst-case reward of a standard non-robust optimal policy is 21% lower than the worst-case reward of the proposed policy. This result shows the robustness of our control policy with respect to distributional errors within Wasserstein distance-based ambiguity sets. Fig. 1 (b) illustrates the sensitivity of the worst-case reward with respect to the radius θ. These local behaviors of the value function are consistent with the global behaviors. Therefore, the proposed sensitivity is useful to select θ without solving the distributionally robust MDP problems for too many θ's. Fig. 2 shows the corresponding indoor temperature distributions controlled by the proposed and standard methods, respectively, when $\theta = 0.1$. Even with a worst-case probability distribution of $(\mathbf{p}_s, \mathbf{r}_s)$, the proposed method can drive the indoor temperature to the most desirable value, $s^* = 20.5$ (°C), while the standard approach cannot.

V. Conclusion

For MDPs, we have proposed several convex optimization-based tools to construct and analyze optimal control policies that are robust against errors in the joint distribution of reward and transition probability vectors; the Wasserstein metric has been employed to measure the error from a nominal distribution. It is to be emphasized that Kantorovich's convex relaxation method and duality principle greatly benefit the proposed tools. This letter could be extended in several ways, such as (*i*) adding a risk constraint to systematically penalize undesirable system behaviors, (*ii*) employing both moment- and Wasserstein distance-based constraints to characterize ambiguity sets in a detailed manner, and (*iii*) developing a scalable numerical method, for example by using approximate dynamic programming.

References

[1] H. Scarf, K. J. Arrow, and S. Karlin, "A min-max solution of an inventory problem," *Studies in the Mathematical Theory of Inventory and Production*. Stanford, CA, USA: Stanford Univ. Press, 1958, pp. 201–209.

[2] F. Miao *et al.*, "Data-driven distributionally robust vehicle balancing using dynamic region partitions," in *Proc. 8th ACM/IEEE Int. Conf. Cyber-Phys. Syst.*, Pittsburgh, PA, USA, 2017, pp. 261–271.

[3] A. Nilim and L. El Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," *Oper. Res.*, vol. 53, no. 5, pp. 780–798, 2005.

[4] G. N. Iyengar, "Robust dynamic programming," *Math. Oper. Res.*, vol. 30, no. 2, pp. 257–280, 2005.

[5] E. Delage and S. Mannor, "Percentile optimization for Markov decision processes with parameter uncertainty," *Oper. Res.*, vol. 58, no. 1, pp. 203–213, 2010.

[6] S. Mannor, O. Mebel, and H. Xu, "Robust MDPs with *k*-rectangular uncertainty," *Math. Oper. Res.*, vol. 41, no. 4, pp. 1484–1509, 2016.

[7] P. Yu and H. Xu, "Distributionally robust counterpart in Markov decision processes," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2538–2543, Sep. 2016.

[8] H. Xu and S. Mannor, "Distributionally robust Markov decision processes," *Math. Oper. Res.*, vol. 37, no. 2, pp. 288–300, 2012.

[9] B. P. G. Van Parys, D. Kuhn, P. J. Goulart, and M. Morari, "Distributionally robust control of constrained stochastic systems," *IEEE Trans. Autom. Control*, vol. 61, no. 2, pp. 430–442, Feb. 2016.

[10] I. Yang, "A dynamic game approach to distributionally robust safety specifications for stochastic systems," *arXiv preprint arXiv:1701.06260*, 2017.

[11] L. N. Vasershtein, "Markov processes on a countable product space, describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–73, 1969.

[12] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *arXiv preprint arXiv:1505.05116*, 2015.

[13] C. Zhao and Y. Guan, "Data-driven risk-averse stochastic optimization with Wasserstein metric," *Available on Optimization Online*, 2015.

[14] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[15] C. Villani, *Topics in Optimal Transportation*, vol. 58. Providence, RI, USA: Amer. Math. Soc., 2003.

[16] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Sci., 1996.

[17] J. Dupačová, "The minimax approach to stochastic programming and an illustrative application," *Stochastics*, vol. 20, no. 1, pp. 73–88, 1987.

[18] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Oper. Res.*, vol. 58, no. 3, pp. 595–612, 2010.

[19] S. Zymler, D. Kuhn, and B. Rustem, "Distributionally robust joint chance constraints with second-order moment information," *Math. Program. A*, vol. 137, no. 1, pp. 167–198, 2013.

[20] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Oper. Res.*, vol. 62, no. 6, pp. 1358–1376, 2014.

[21] G. Bayraksan and D. K. Love, "Data-driven stochastic programming using phi-divergences," in *Proc. Tutorials Oper. Res.*, 2015, pp. 1–19.

[22] A. Ben-Tal, D. D. Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Manag. Sci.*, vol. 59, no. 2, pp. 341–357, 2013.

[23] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Math. Program. A*, vol. 158, no. 1, pp. 291–327, 2016.

[24] E. Erdoğan and G. Iyengar, "Ambiguous chance constrained problems and robust optimization," *Math. Program. B*, vol. 107, no. 1, pp. 37–61, 2006.

[25] M. Breton, A. Alj, and A. Haurie, "Sequential Stackelberg equilibria in two-person games," *J. Optim. Theory Appl.*, vol. 59, no. 1, pp. 71–97, 1988.

[26] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 2014.

[27] P. Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.