Stochastic and Distributionally Robust Load Ensemble Control

Ali Hassan, Robert Mieth, Deepjyoti Deka and Yury Dvorkin

***Abstract*—Demand response (DR) programs aim to engage distributed demand-side resources in providing ancillary services for electric power systems. Previously, aggregated thermostatically controlled loads (TCLs) have been demonstrated as a technically viable and economically valuable provider of such services that can effectively compete with conventional generation resources in reducing load peaks and smoothing demand fluctuations. Yet, to provide these services at scale, a large number of TCLs must be accurately aggregated and operated in sync. This paper describes a Markov Decision Process (MDP) that aggregates and models an ensemble of TCLs. Using the MDP framework, we propose to internalize the exogenous uncertain dynamics of TCLs by means of stochastic and distributionally robust optimization. First, under mild assumptions on the underlying uncertainty, we derive analytical stochastic and distributionally robust control policies for dispatching a given TCL ensemble. Second, we further relax these mild assumptions to allow for a more delicate treatment of uncertainty, which leads to distributionally robust MDP formulations with moment- and Wasserstein-based ambiguity sets that can be efficiently solved numerically. The case study compares the analytical and numerical control policies using a simulated ensemble of 1,000 air conditioners.**

***Index Terms*—Markov Decision Process (MDP), Linearly Solvable MDP, Distributionally Robust MDP, Thermostatically Controlled Loads, Uncertainty**

I. Introduction

Thermal inertia of cooling and heating systems enables temporarily adjusting power consumption of thermostatically controlled loads (TCLs) without compromising their primary functions [1], [2]. In the presence of constantly growing volatility and uncertainty of nodal power injections in electric power distribution systems caused by the integration of distributed energy resources (DERs), thermal flexibility of TCLs is a valuable control resource, [1]. The ongoing expansion of grid-edge communication infrastructure also allows for designing demand response (DR) programs that enroll distributed small-scale flexible loads to provide various grid support services, both at the transmission and distribution levels. The Federal Energy Regulatory Commission (FERC) reports an increasing trend of DR program participation in the wholesale markets with a growth of 3% from 2016 to 2017, to a total of 27,541 MW [3]. To a large extent, this participation is enabled by aggregators that operate a large portfolio of similar devices [4] (called an ensemble) and act as mediators between grid operating entities, e.g. distribution system operators (DSOs), and individual flexible loads. The efficiency of these DR programs depends on the ability of aggregators to accurately model and control their ensembles.

TCLs, such as air conditioners, refrigerators or electric heaters, have a cycling pattern of energy consumption, i.e. they switch between on and off states given some user-defined thresholds (e.g. preferred temperature bands). This property allows to model TCL ensembles of an unlimited, or sufficiently large, size as a discrete-time, discrete-space Markov Process (MP) with relatively high accuracy. Their power consumption can then be optimized using the Markov Decision Process (MDP) framework [5]–[13]. The MP approach exploits the on/off switching behavior of TCLs and discretizes the ensemble dynamics into a finite number of states with each possible transition between these states characterized by a state-dependent probability. By capturing these transitions and their probabilities, the MP characterizes the interplay between the TCL temperature settings and electrical consumption based on external parameters (e.g. quality of refrigerator insulation, volume of air-conditioned space).

In [5], the authors show that the necessary parameters to construct such a MP representation can be obtained either from TCL electrical measurements or system temperature observations. The MP in [5] then employs a model predictive control strategy to achieve a desired consumption trajectory of the ensemble, thus allowing for dispatching TCLs like a virtual energy storage device. Similarly, [14], [15] developed methods to represent and dispatch TCL ensembles as virtual storage devices for providing regulation reserve. In the context of DR aggregators, the desired load trajectory is the optimal trade-off between increasing the payoff of the aggregator and reducing comfort levels of TCL users, e.g. discomfort caused by deviations from their temperature settings. By penalizing deviations from user-defined TCL settings, the MDP in [6], [7] provides a tractable description of the TCL optimization by leveraging dynamic programming. MDP-based DR frameworks similar to [5]–[7] can also accommodate network constraints to account for AC power flow and voltage limits in the distribution system [8], [9], as well as to mitigate the uncertainty of PV generation resources [10]. In addition, [11] consider the effect of fluctuating electricity prices on various types of controllable loads and derive a price-taking control strategy. Unlike [5]–[11], [14], [15], which assume that TCLs are operated in a centralize manner, [12], [13] develop a decentralized Markovian control strategy for an individual TCL resource to provide ancillary services to the power grid.

While [5]–[13] demonstrate the usefulness of the MDP framework for dispatching TCL ensembles, they assume per-

This work at NYU was supported in part by the National Science Foundation (NSF) under Award EECS-1847285 and in part by the U.S. Department of Energy under Award DE-AC52-07NA27344.

fect knowledge of the ensemble transitions and their probabilities. In practice, however, these model parameters are unknown and must be inferred from historical data. As available data on TCL ensembles is finite and potentially noisy, the true values of these model parameters remain unknown. This paper robustifies the MDP-based optimization of a risk-averse DR aggregator against uncertainty in the transition probabilities, thus generalizing the MDP models in [6]–[8], [10]. Leveraging methods of stochastic and distributionally robust optimization, we derive analytical and numerical methods to endogenously model uncertain transition probabilities and explore their potential effects on the optimal dispatch of TCL ensembles.

Parameter uncertainty arising from the inability to accurately estimate transition probabilities of the MP has been shown to significantly distort the outcomes of MDP solutions [16]. The most common methods to overcome this caveat include percentile criteria [17], Kullback-Leibler divergence bounds [18], nested uncertainty sets [19] or confidence regions using historical MDP performance metrics [20]. This paper exploits an alternative approach and aims to internalize statistical information about the uncertainty on transition probabilities into the MDP optimization. Specifically, we explore how a mildly restrictive assumption enables a reformulation of the MDP optimization for TCLs as a linearly-solvable MDP (LS-MDP) [21]. Using this LS-MDP framework and building on the previous work in [7], [8], [10], this paper accounts for the transition probability uncertainty in the MDP optimization under different statistical assumptions summarized in Table I. First, we use stochastic and distributionally robust optimization to derive analytical (closed-loop) control policies for the TCL ensembles under the assumption that the transition probability uncertainty is normally distributed, either with known or ambiguous distribution parameters. However, this assumption may still lead to unnecessarily erroneous TCL dispatch decisions. Second, we overcome the need for the normally distributed assumption, by introducing a moment-based ambiguity set into the MDP optimization that does not assume any distribution and only requires knowledge about first- and second-order moments. Although this approach does not result in a closed-form optimal control policy, we demonstrate that the MDP optimization under these assumptions can be solved efficiently with off-the-shelf solvers. To overcome the requirement on accurately computing the moments, we introduce a Wasserstein probability distance, [22], [23], in the distributionally robust MDP optimization and derive a computationally tractable reformulation. Unlike the moment-based approach, the Wasserstein allows to capture all distributions within a pre-defined radius from a given nominal distribution, which can be drawn from empirical data, thus reducing data requirements needed to obtain a distributionally robust solution. Furthermore, the value of this radius can be used by decision-makers as a tuning parameter that allows for adjusting the solution conservatism. To demonstrate and compare the performance of the presented analytical and numerical approaches, we conduct comprehensive numerical experiments on a TCL ensemble consisting of air conditioners.

Table I. OVERVIEW OF THE EXISTING AND PROPOSED METHODS

Method	Eq.	Uncertainty on transition probability	Solution
Previous work, [7], [8], [10]	(1)	None	Analytical
Stochastic	(7)	Normally distributed	Analytical
Distributionally robust	(13)	Normally distributed with ambiguous parameters	
Moment-based distributionally robust	(20)	Any distribution with constraints on moments	Numerical
Wasserstein-based distributionally robust	(23)	Any distribution within a fixed distance of empirical distribution	

II. MDP FOR TCL ENSEMBLES

Building on our prior work in [7], [8], [10], we represent a homogeneous ensemble of sufficiently many TCLs as a discrete-time, discrete-space MDP. From the perspective of the DR aggregator, the optimization problem for operating the TCL ensemble is:

$$\min_{\rho,\mathcal{P}_t} \mathbb{E}_\rho \sum_{t\in\mathcal{T}}\sum_{\alpha\in\mathcal{A}} \big(-U^\alpha_{t+1} + \sum_{\beta\in\mathcal{A}} \gamma \log \frac{\mathcal{P}^{\alpha\beta}_t}{\overline{\mathcal{P}}^{\alpha\beta}} \big) \quad (1a)$$

$$\text{s.t.} \quad \rho^\alpha_{t+1} = \sum_{\beta\in\mathcal{A}} \mathcal{P}^{\alpha\beta}_t \rho^\beta_t, \qquad \forall \alpha \in \mathcal{A}, t \in \mathcal{T} \quad (1b)$$

$$\sum_{\alpha\in\mathcal{A}} \mathcal{P}^{\alpha\beta}_t = 1, \qquad \forall \beta \in \mathcal{A}, t \in \mathcal{T} \quad (1c)$$

where $\rho \in \mathbb{R}^n$, $n = |\mathcal{A}|$, is a vector with entries $\rho^\alpha_{t+1} \geq 0$ and $\rho^\beta_t \geq 0$ representing the probabilities that the TCL ensemble is in states $\alpha, \beta \in \mathcal{A}$ at times $t+1$ and t, respectively, $\mathcal{A}$ is the set of all possible states, and operator $\mathbb{E}_\rho$ denotes the expectation over ρ. Set $\mathcal{A}$ is obtained by discretizing the range of aggregated power consumption of the ensemble given the operating range of each TCL [7]. Probabilities ρ^α_{t+1} and ρ^β_t are related via the transition probability matrix $\mathcal{P}_t \in \mathbb{R}^{nxn}$, with $n = |\mathcal{A}|$, and where entry $\mathcal{P}^{\alpha\beta}_t$ of matrix $\mathcal{P}_t$ characterizes the probability of the transition of the TCL ensemble from state β at time t to state α at time $t+1$. Note that the TCL ensemble can also remain in the same state such that $\alpha = \beta$. On the other hand, matrix $\overline{\mathcal{P}} \in \mathbb{R}^{nxn}$ with entries $\overline{\mathcal{P}}^{\alpha\beta}$ represents the default transition probability, i.e. the steady state behavior of the ensemble without any control actions of the aggregator. Additionally, internal control actions such as user-defined settings and their on-demand adjustments can still be applied to the individual TCLs in the ensemble, which will modify and will be reflected in default transitions and the probability matrix. (The inability to perfectly forecast these internal control actions introduce the uncertainty that we deal with in Sections III-IV.) In the following, we treat the vector ρ and matrix $\mathcal{P}_t$ as decision variables, which can be achieved by suitable TCL control actions [5]. In contrast, entries $\overline{\mathcal{P}}^{\alpha\beta}$ of matrix $\overline{\mathcal{P}}$ are treated as parameters of the MDP optimization in (1). Although matrix $\overline{\mathcal{P}}$ is modeled as time-independent, unlike $\mathcal{P}_t$, this modeling choice can be revisited, if sufficient historical data about the TCL ensemble is available. As more empirical data on the TCL dispatch is collected over time,

the more temporal fidelity can be achieved in representing default transitions. All methods to account for the uncertainty presented below will hold if $\overline{\mathcal{P}}$ is modeled as time-dependent.

Eq. (1a) is the objective function of the aggregator that operates the TCL ensemble and tries to maximize its expected utility U^{α}_{t+1} at future state α at time $t+1$ and to minimize the discomfort cost of the TCL ensemble, which is modeled as the logarithmic difference between the uncontrolled transitions of the TCL ensemble ($\overline{\mathcal{P}}^{\alpha\beta}$) and the resulting transition probabilities due to the control decisions of the aggregator ($\mathcal{P}^{\alpha\beta}_t$). [1] This discomfort cost in the second term of (1a) can be interpreted as the Kullback-Leibler (KL) divergence weighed by cost penalty γ, [24]. The KL divergence is widely used for modeling discrepancies in discrete- and continuous-time series, [25], and makes it possible to derive closed-form optimal control policies. Parameter γ can influence the KL divergence and thus encourage or discourage deviations from the default behavior of the TCL ensemble. Furthermore, if $\overline{\mathcal{P}}^{\alpha\beta} = 0$, i.e. a transition from state β to α has not been observed in the past, the model in (1) restricts $\mathcal{P}^{\alpha\beta}_t = 0$ and excludes such transitions when optimizing it for the rest of the values. Eq. (1b) describes the temporal evolution of the TCL ensemble from time t to $t+1$ over time horizon $\mathcal{T}$. Eq. (1c) imposes the integrality constraint on the transition decisions optimized by the aggregator such that their total probability is equal to one. After solving (1), the active power (p_t) consumed by the TCL ensemble can be computed using decisions ρ^{β}_t and rated active power $p^{\beta,rated}$ at each state as $p_t = \sum_{\beta\in\mathcal{A}} p^{\beta}\rho^{\beta,rated}_t, \forall t \in \mathcal{T}$. Since (1) is formulated for a discrete-time MP, the resulting dispatch does not capture power fluctuations between discrete time instances. However, since the TCL ensemble is assumed to be sufficiently large, random fluctuations of TCL loads neutralize one another at the ensemble level, [26]. Furthermore, the residual effects of such fluctuations between discrete time instances can be mitigated if one uses a more fine-grained temporal resolution. However, the latter may increase computing times.

Our prior work in [8], [10] shows that the optimization in (1) is a LS-MDP as introduced by [21]. The LS-MDP has no explicit actions, is controlled by modifying a predefined (uncontrolled) probability distribution over subsequent states as modeled by decisions $\mathcal{P}^{\alpha\beta}_t$. The optimal policy obtained from (1) is a next-state distribution, which minimizes the accumulated state costs of the agent traversing state space $\mathcal{A}$, while minimizing the divergence cost between the controlled ($\mathcal{P}^{\alpha\beta}_t$) and uncontrolled ($\overline{\mathcal{P}}^{\alpha\beta}$) probability distributions. This optimal policy can be computed as:

Theorem 1. Let (1) model a TCL ensemble as a LS-MDP. Then the optimal control policy is:

$$\mathcal{P}^{\alpha\beta}_t = \frac{\overline{\mathcal{P}}^{\alpha\beta} z^{\alpha}_{t+1}}{\sum_{\alpha\in\mathcal{A}} \overline{\mathcal{P}}^{\alpha\beta} z^{\alpha}_{t+1}}, \tag{2}$$

where $z^{\alpha}_{t+1} = \exp(-\varphi^{\alpha}_{t+1}/\gamma)$ and value function φ^{α}_{t+1} is defined as $\varphi^{\alpha}_{t+1} = -U^{\alpha}_{t+1} - \gamma \log\big(\sum_{\upsilon\in\mathcal{A}} \exp\big(\frac{-\varphi^{\upsilon}_{t+2}}{\gamma}\big)\overline{\mathcal{P}}^{\upsilon\alpha}\big)$, where $\upsilon \in \mathcal{A}$ is a state at time $t+2$.

Proof. See proof in Appendix A. □

Theorem 1 implies that computing the optimal control policy depends on the uncontrolled transition probability ($\overline{\mathcal{P}}^{\alpha\beta}$) and the value function of the next state (φ^{α}_{t+1}). However, this requires the default transition probabilities to be perfectly known, which does not hold in real-world applications, where the TCL ensemble is subject to unknown external influences and uncertain human behavior. We model this parameter uncertainty by representing default transition probabilities $\overline{\mathcal{P}}^{\alpha\beta}$ as random variables $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$, indicated by the **bold** font. As summarized in Table I, we derive and study methods to internalize $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$ in the optimal MDP control policy using different assumptions and statistical information on $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$.

Remark. Although the MDP in (1) is developed for a homogenous TCL ensemble, it can be extended to modeling heterogenous TCL ensembles. For instance, one can classify TCL loads in a given heterogenous ensemble and represent it as a set of homogeneous subensembles. Then, each subensemble can be operated separately using the proposed MDP framework. Similarly, the models proposed in Sections III and IV can be extended to operating heterogenous TCL ensembles.

III. Analytical Control Policies

The standard MDP formulation in (1) allows the derivation of a closed-form optimal control policy as shown by Theorem 1. The goal of this section is to show that this useful property can be maintained if $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$ is normally distributed.

A. Stochastic Formulation

Assume that $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$ follows a normal distribution with mean $\overline{\mathcal{P}}^{\alpha\beta}$ and variance σ^2, i.e. $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \sim N(\overline{\mathcal{P}}^{\alpha\beta}, \sigma^2)$. The mean and variance can be calculated from a set of N historical observations of $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$ that can be retrieved by the aggregator from operating data of a given TCL ensemble[2]. We denote this set of observations as $\{\overline{\mathcal{P}}^{\alpha\beta}_{j,obs}\}_{j\in N}$ and use it to infer distribution parameters such as empirical mean ($\overline{\mathcal{P}}^{\alpha\beta}$) and variance ($\sigma^2$) as follows:

$$\overline{\mathcal{P}}^{\alpha\beta} = \frac{1}{N}\sum_{j\in N} \overline{\mathcal{P}}^{\alpha\beta}_{j,obs}, \;\; \sigma^2 = \frac{1}{N-1}\sum_{j\in N} (\overline{\mathcal{P}}^{\alpha\beta}_{j,obs} - \overline{\mathcal{P}}^{\alpha\beta})^2 \tag{3}$$

Then, we reformulate (1) as:

$$\min_{\rho,\mathcal{P}} \; O^E := \mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}} \mathbb{E}_{\rho} \sum_{t\in\mathcal{T}}\sum_{\alpha\in\mathcal{A}} \Bigg(-U^{\alpha}_{t+1} + \sum_{\beta\in\mathcal{A}} \gamma \log \frac{\mathcal{P}^{\alpha\beta}_t}{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}} \Bigg) \tag{4a}$$

$$\text{s.t.} \quad \text{Eq. (1b)} - \text{(1c)}, \tag{4b}$$

[1] The discomfort cost of TCLs can be interpreted as a change in their temperature settings from user-defined comfort/convenience levels, e.g. for freezers, air-conditioners, hot-water tanks, heat pumps, and swimming pool pumps.

[2] We ensure $\sum_{\alpha\in\mathcal{A}} \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} = 1$. In other words, the probability of moving from present state β to all possible next states α is equal to one.

where $\mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}}$ denotes the expectation over $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$ and $\mathbb{E}_\rho$ is identical to (1a). Eq. (4a) can further be simplified as:

$$O^E = \mathbb{E}_\rho \sum_{t\in\mathcal{T}}\sum_{\alpha\in\mathcal{A}}\Big\{-U^\alpha_{t+1} + \gamma \sum_{\beta\in\mathcal{A}}\big(\log \mathcal{P}^{\alpha\beta}_t - \mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}}[\log \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}]\big)\Big\} \quad (5)$$

where the last term can be approximated by the second-order Taylor expansion as, [27, Eq. (17)]:

$$\begin{aligned}\mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}}[\log(\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta})] &\approx \log \mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}}[\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}] - \frac{\mathrm{Var}(\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta})}{2(\mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}}[\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}])^2} \\ &= \log(\overline{\mathcal{P}}^{\alpha\beta}) - \frac{\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}. \end{aligned} \quad (6)$$

Given (6), the optimization in (4) is rewritten as:

$$\begin{aligned}\min_{\rho,\mathcal{P}}\ O^E := \mathbb{E}_\rho \sum_{t\in\mathcal{T}}\sum_{\alpha\in\mathcal{A}}\Big\{ &- U^\alpha_{t+1} + \gamma \sum_{\beta\in\mathcal{A}}\Big(\log \frac{\mathcal{P}^{\alpha\beta}_t}{\overline{\mathcal{P}}^{\alpha\beta}} \\ &+ \frac{\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\Big)\Big\} \end{aligned} \quad (7a)$$

$$\text{s.t.} \quad \text{Eq. (1b)} - \text{(1c)} \quad (7b)$$

Given the stochastic formulation in (7), we prove:

Theorem 2. Let (7) model a TCL ensemble as a LS-MDP with uncertain transition probabilities defined as $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \sim N(\overline{\mathcal{P}}^{\alpha\beta}, \sigma^2)$. Then the optimal control policy is:

$$\mathcal{P}^E_t := \mathcal{P}^{\alpha\beta}_t = \frac{\overline{\mathcal{P}}^{\alpha\beta} z^\alpha_{t+1} \exp\big(\frac{-\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\big)}{\sum_\alpha \overline{\mathcal{P}}^{\alpha\beta} z^\alpha_{t+1} \exp\big(\frac{-\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\big)}, \quad (8)$$

where $z^\alpha_{t+1} = \exp(-\varphi^\alpha_{t+1}/\gamma)$ and value function φ^α_{t+1} is defined as $\varphi^\alpha_{t+1} = -U^\alpha_{t+1} - \gamma \log\big(\sum_{\upsilon\in\mathcal{A}} \exp\big(\frac{-\varphi^\upsilon_{t+2}}{\gamma}\big) \overline{\mathcal{P}}^{\upsilon\alpha} \exp\big(\frac{-\sigma^2}{2(\overline{\mathcal{P}}^{\upsilon\alpha})^2}\big)\big)$, where $\upsilon \in \mathcal{A}$ is a state at time $t+2$.

Proof. See proof in Appendix A. □

Similarly to Theorem 1, the optimal control policy obtained from Theorem 2 depends on the mean values of uncontrolled transition probabilities ($\overline{\mathcal{P}}^{\alpha\beta}$), the next-state value function (φ^a_{t+1}) and variance (σ^2). However, term $\exp\Big(\frac{-\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\Big)$ distinguishes the control policy in Theorem 2 from Theorem 1 and internalizes the uncertainty on uncontrolled transition probabilities into the optimal control policy. Hence, the stochastic solution in Theorem 2 is anticipated to improve the optimal control policy formulated in Theorem 1 for an average performance of the TCL ensemble. However, Theorem 2 still exploits the assumption that parameters of the uncertainty distribution, i.e. $\overline{\mathcal{P}}^{\alpha\beta}$ and σ^2, are perfectly known.

B. Distributionally Robust Formulation

To internalize potential parameter misestimation due to the finite number of available observations, we leverage distributionally robust optimization that allows for modeling the inferred distribution parameters via an ambiguity set. In this setting, the objective of the DR aggregator is to maximize their expected performance under the worst-case distribution of $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$ drawn from a given ambiguity set denoted as $\mathbb{D}$:

$$\begin{aligned}\min_{\rho,\mathcal{P}}\ O^{WC} := \sup_{\overline{\boldsymbol{\mathcal{P}}}\in\mathbb{D}} \mathbb{E}_\rho \sum_{t\in\mathcal{T}}\sum_{\alpha\in\mathcal{A}}\Big\{ &- U^\alpha_{t+1} \\ &+ \gamma \sum_{\beta\in\mathcal{A}}\Big(\log \frac{\mathcal{P}^{\alpha\beta}_t}{\overline{\mathcal{P}}^{\alpha\beta}} + \frac{\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\Big)\Big\} \end{aligned} \quad (9a)$$

$$\text{s.t.} \quad \text{Eq. (1b)} - \text{(1c)}, \quad (9b)$$

The ambiguity set in (9) is defined as $\mathbb{D} = [\underline{\Gamma} \le \overline{\mathcal{P}}^{\alpha\beta} \le \overline{\Gamma}, \underline{\hat{\zeta}} \le \sigma^2 \le \overline{\hat{\zeta}}]$, where $\underline{\Gamma}$, $\overline{\Gamma}$, $\underline{\hat{\zeta}}$ and $\overline{\hat{\zeta}}$ are confidence bounds on the empirical mean and variance. Since $\overline{\mathcal{P}}^{\alpha\beta}$ and σ^2 can be respectively modeled by t- and Chi-Square ($\mathcal{X}^2$) distributions [28], we compute these bounds as:

$$\underline{\Gamma} = \overline{\mathcal{P}}^{\alpha\beta} - t_{(1-\varsigma/2)} \frac{\sigma}{\sqrt{N}} \ \text{ and } \ \overline{\Gamma} = \overline{\mathcal{P}}^{\alpha\beta} + t_{(1-\varsigma/2)} \frac{\sigma}{\sqrt{N}}, \quad (10)$$

$$\underline{\hat{\zeta}} = \frac{(N-1)\sigma^2}{\mathcal{X}^2_{(1-\xi)/2}} \ \text{ and } \ \overline{\hat{\zeta}} = \frac{(N-1)\sigma^2}{\mathcal{X}^2_{\xi/2}}, \quad (11)$$

where we denote $t_{(1-\varsigma/2)}$ in (10) as the $(1-\varsigma/2)$-quantile of the t-distribution and $\mathcal{X}^2_\xi$ in (11) as the ξ-quantile of the Chi-Square distribution. Given $\mathbb{D}$, the objective function in (9a) can be reformulated as:

$$\begin{aligned}&\sup_{\overline{\boldsymbol{\mathcal{P}}}\in\mathbb{D}} \sum_{t\in\mathcal{T}}\sum_{\alpha\in\mathcal{A}}\Big\{- U^\alpha_{t+1} + \gamma \sum_{\beta\in\mathcal{A}}\Big(\log \frac{\mathcal{P}^{\alpha\beta}_t}{\overline{\mathcal{P}}^{\alpha\beta}} + \frac{\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\Big)\Big\} \\ &= \sum_{t\in\mathcal{T}}\sum_{\alpha\in\mathcal{A}}\Big\{- U^\alpha_{t+1} + \gamma \sum_{\beta\in\mathcal{A}}\Big(\log \frac{\mathcal{P}^{\alpha\beta}_t}{\underline{\Gamma}} + \frac{\overline{\hat{\zeta}}}{2(\underline{\Gamma})^2}\Big)\Big\}, \end{aligned} \quad (12)$$

leading to the following optimization problem:

$$\begin{aligned}\min_{\rho,\mathcal{P}}\ O^{WC} := \mathbb{E}_\rho \sum_{t\in\mathcal{T}}\sum_{\alpha\in\mathcal{A}}\Big\{ &- U^\alpha_{t+1} + \gamma \sum_{\beta\in\mathcal{A}}\Big(\log \frac{\mathcal{P}^{\alpha\beta}_t}{\underline{\Gamma}} \\ &+ \frac{\overline{\hat{\zeta}}}{2(\underline{\Gamma})^2}\Big)\Big\} \end{aligned} \quad (13a)$$

$$\text{s.t.} \quad \text{Eq. (1b)} - \text{(1c)}. \quad (13b)$$

Given the reformulation of (9) presented in (13), we prove:

Theorem 3. Let (13) model a TCL ensemble as a LS-MDP with $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \sim N(\overline{\mathcal{P}}^{\alpha\beta}, \sigma^2)$ and $\overline{\mathcal{P}}^{\alpha\beta}, \sigma^2 \in \mathbb{D}$, where $\mathbb{D} = [\underline{\Gamma} \le \overline{\mathcal{P}}^{\alpha\beta} \le \overline{\Gamma}, \underline{\hat{\zeta}} \le \sigma^2 \le \overline{\hat{\zeta}}]$. Then the optimal control policy is:

$$\mathcal{P}^{WC}_t := \mathcal{P}^{\alpha\beta}_t = \frac{\underline{\Gamma} z^\alpha_{t+1} \exp\big(\frac{-\overline{\hat{\zeta}}}{2(\underline{\Gamma})^2}\big)}{\sum_\alpha \underline{\Gamma} z^\alpha_{t+1} \exp\big(\frac{-\overline{\hat{\zeta}}}{2(\underline{\Gamma})^2}\big)} \quad (14)$$

where $z^\alpha_{t+1} = \exp(-\varphi^\alpha_{t+1}/\gamma)$ and value function φ^α_{t+1} is defined as $\varphi^\alpha_{t+1} = -U^\alpha_{t+1} - \gamma \log\big(\sum_{\upsilon\in\mathcal{A}} \exp\big(\frac{-\varphi^\upsilon_{t+2}}{\gamma}\big) \underline{\Gamma} \exp\big(\frac{-\hat{\zeta}}{2(\underline{\Gamma})^2}\big)\big)$, where $\upsilon \in \mathcal{A}$ is a state at time $t+2$.

Proof. See proof in Appendix A. □

Similarly to Theorems 1 and 2, Theorem 3 computes the optimal control policy using the mean values of the default transition probabilities ($\overline{\mathcal{P}}^{\alpha\beta}$) and the next-state value function (φ^α_{t+1}). However, it additionally internalizes the information about set $\mathbb{D}$ and immunize the optimal control policy for the

worst-case realization of distribution parameters drawn from this set. This overcomes the need to perfectly know distribution parameters as in Theorem 2, thus improving the goodness of fit between the LS-MDP model and empirical data.

C. Hybrid Model

Relative to the stochastic formulation in (7), the distributional robustness of (13) imposes additional conservatism on the optimal control policy, which may lead to a greater solution cost. To trade-off the robustness and cost performance of the optimal policy, we seek the hybrid formulation that can weigh the stochastic and distributionally robust formulations via parameter η:

$$\min_{\rho,\mathcal{P}} \ (1-\eta)O^{WC} + \eta O^{E} \tag{15a}$$

$$\text{s.t.} \quad \text{Eq. (1b)} - \text{(1c)} \tag{15b}$$

where $0 \leq \eta \leq 1$.

Theorem 4. Let (15) model a TCL ensemble as a LS-MDP with uncertainty defined as $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \sim N(\overline{\mathcal{P}}^{\alpha\beta}, \underline{\sigma}^2)$ and $\overline{\mathcal{P}}^{\alpha\beta}$,$\sigma^2 \in \mathbb{D}$, where $\mathbb{D} = [\underline{\Gamma} \leq \overline{\mathcal{P}}^{\alpha\beta} \leq \overline{\Gamma}, \hat{\underline{\zeta}} \leq \sigma^2 \leq \hat{\overline{\zeta}}]$. Then the optimal control policy is given as:

$$\mathcal{P}_t^{\alpha\beta} = (1-\eta)\mathcal{P}_t^{WC} + \eta \mathcal{P}_t^{E} \tag{16}$$

where $0 \leq \eta \leq 1$ is a parameter characterizing risk tolerance of the aggregator and $\mathcal{P}_t^{E}$ and $\mathcal{P}_t^{WC}$ are given by (8) and (14).

Proof. See proof in Appendix A. □

Theorem 4 yields the optimal control policy that balances the stochastic and distributionally robust models weighted by parameter η, which can be set by the DR aggregator based on its risk tolerance.

IV. Numerical Control Policies

The analytical control policies derived in the previous Section III assume that $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$ is normally distributed, even if distribution parameters are not precisely known and drawn from the ambiguity set. However, these assumptions may still limit the performance and applicability of the analytical policies. This caveat motivates a further investigation of methods that allow for more generic control policies.

A. Moment-based Ambiguity Set

Instead of assuming a specific (e.g. normal) uncertainty distribution, we define $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$ solely in terms of its statistical moments (e.g. mean and variance). In other words, this approach achieves distributional robustness by defining an ambiguity set that captures all distributions with statistical moments satisfying given confidence parameters. Hence, we redefine uncertainty set $\mathbb{D}$:

$$\begin{aligned}\mathbb{D} := \{\mathbb{P} \in \mathcal{M}(\mathbb{R}) | \mathbb{P}(W) = 1 : (\nu),\\ -b \leq \mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \sim \mu}[\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}] - m \leq b : (\underline{\lambda}, \overline{\lambda}),\\ [\mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \sim \mu}(\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} - m)^2] \leq c\sigma^2 : (\Lambda)\},\end{aligned} \tag{17}$$

where $\mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \sim \mathbb{P}}$ is the expectation over empirical probability distribution $\mathbb{P}$ supported by samples $\{\overline{\mathcal{P}}_y^{\alpha\beta}\}_{y \in N}$, $\mathcal{M}$ is the set of all distributions, W is the support set, and m and σ^2 are the nominal mean and variance with confidence parameters b and c. Given the nominal values and confidence parameters, the uncertainty set in (17) allows for the worst-case mean and variance be drawn from a range of values. Note that in (17) we introduce dual variables $\nu, \underline{\lambda}, \overline{\lambda}$, and Λ for each constraint, which are given after a colon. Given the ambiguity set in (17), we define the following optimization problem:

$$\min_{\rho,\mathcal{P}} \sup_{\mathbb{P} \in \mathbb{D}} \mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \sim \mathbb{P}} \mathbb{E}_{\rho} \sum_{t \in \mathcal{T}} \sum_{\alpha \in \mathcal{A}} \left(- U_{t+1}^{\alpha} + \gamma \sum_{\beta \in \mathcal{A}} \log \frac{\mathcal{P}_t^{\alpha\beta}}{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}} \right) \tag{18a}$$

$$\text{s.t.} \quad \text{Eq. (1b)} - \text{(1c)}. \tag{18b}$$

Solving (18) is challenging because the optimization is performed over infinite dimensional set $\mathbb{D}$. To the best of our knowledge, such problems cannot be solved analytically and there are also no efficient computational tools [29]. However, one way to tackle such problems is to leverage convex duality theory that transforms the original problem over an infinite dimensional set into a dual problem over finite dimensional Lagrange multipliers with the same value as the original problem [30]–[32]. The duality approach in an infinite dimensional setting is developed by Rockafellar in [30] and is based on pairing locally convex topological vector spaces. The requirement of the existence of a feasible interior point (Karush-Kuhn-Tucker point) for the implicit constraint set is relaxed to require only continuity of the optimal value function. After transforming the original problem to its dual form, we can use finite optimization computational tools to obtain a solution. Therefore, we take the dual of the inner maximization problem and reformulate the objective function (18a) as follows:

$$\begin{aligned}\min_{\underline{\lambda},\overline{\lambda},\Lambda,\nu} \mathbb{E}_{\rho} \sum_{t \in \mathcal{T}} \sum_{\alpha \in \mathcal{A}} \Bigg\{ - U_{t+1}^{\alpha} + \gamma \sum_{\beta \in \mathcal{A}} \Big[\log \mathcal{P}_t^{\alpha\beta}\\ + (b-m)\underline{\lambda} + (b+m)\overline{\lambda} + c\sigma^2\Lambda + \nu \Big] \Bigg\}\end{aligned} \tag{19a}$$

s.t.

$$(\overline{\lambda} - \underline{\lambda})\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} + \Lambda(\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} - m)^2 + \nu \geq -\log \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}, \ \forall \overline{\mathcal{P}}^{\alpha\beta} \in W, \tag{19b}$$

where $\{\underline{\lambda}, \overline{\lambda}, \Lambda \geq 0; \ \nu : \text{free}\}$ are dual variables defined for the constraints in ambiguity set $\mathbb{D}$ given by (17). Eq. (19) represents an upper bound of the inner maximization in (18) because (18a) essentially maximizes over a convex function ($\sup - \log \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$). By substituting (19) in (18), we obtain the following single-level optimization problem:

$$\begin{aligned}\min_{\rho,\mathcal{P},\underline{\lambda},\overline{\lambda},\Lambda,\nu} \mathbb{E}_{\rho} \sum_{t \in \mathcal{T}} \sum_{\alpha \in \mathcal{A}} \Bigg\{ - U_{t+1}^{\alpha} + \gamma \sum_{\beta \in \mathcal{A}} \Big[\log \mathcal{P}_t^{\alpha\beta}\\ + (b-m)\underline{\lambda} + (b+m)\overline{\lambda} + c\sigma^2\Lambda + \nu \Big] \Bigg\}\end{aligned} \tag{20a}$$

$$\text{s.t.} \quad \text{Eq. (1b)} - \text{(1c)} \tag{20b}$$

$$\text{Eq. (19b)}. \tag{20c}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2020.2992268, IEEE Transactions on Power Systems

The optimization problem in (20) can be solved numerically with off-the-shelf solvers by discretizing W in (19b). Note that relative to the analytical control policies developed in Section III, (20) yields a numerical solution, yet with optimality guarantees. Although this numerical solution is less generalizable than the analytical solutions, it is obtained under less restrictive assumptions on the underlying uncertainty, which is more suitable for practical needs and allows one to avoid unnecessary conservatism of the optimal solution.

B. Wasserstein-based Ambiguity Set

Although the moment-based ambiguity set in (17) avoids assuming a particular distribution, it still restricts the first- and second-order moments within given ranges determined by the confidence parameters, which is shown to produce overly conservative solutions for certain problems [23]. Hence, to alleviate the need to invoke these restrictions, we define an ambiguity set using the Wasserstein metric, which makes it possible to immunize the optimal solution against any distribution that lies within fixed radius $\psi > 0$ around a given nominal distribution. Accordingly, we formulate this ambiguity set as:

$$\mathcal{C}_\tau := \{\mathbb{P} \in \mathcal{M} : W_p(\mathbb{P}, \hat{\mathbb{P}}) \leq \psi\}, \tag{21}$$

where W_p is the Wasserstein metric of order p evaluating the distance between distribution $\mathbb{P}$ and nominal distribution $\hat{\mathbb{P}}$. Given empirical distribution $\overline{\boldsymbol{\mathcal{P}}}_t^{\alpha\beta}$ based on observations $\{\overline{\mathcal{P}}_y^{\alpha\beta}\}_{y\in N}$, the nominal distribution in (21) can be defined as $\hat{\mathbb{P}} = \frac{1}{N}\sum_{y\in N}\delta_{\overline{\mathcal{P}}_y^{\alpha\beta}}$, where $\delta_{\overline{\mathcal{P}}_y^{\alpha\beta}}$ is a Dirac distribution for $\overline{\boldsymbol{\mathcal{P}}}_y^{\alpha\beta}$. Hence, the Wasserstein distance between distributions $\mathbb{P}$ and $\hat{\mathbb{P}}$ defines the minimum cost of redistributing mass from $\mathbb{P}$ to $\hat{\mathbb{P}}$. Hence, using (21), we can reformulate the distributionally robust objective function as follows:

$$\min_{\rho,\mathcal{P}} \sup_{\mathbb{P}\in\mathcal{C}_\tau} \mathbb{E}_{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}} \mathbb{E}_\rho \sum_{t\in\mathcal{T}} \Bigg(\sum_{\alpha\in\mathcal{A}} -U_{t+1}^{\alpha} + \gamma \sum_{\alpha\in\mathcal{A}} \sum_{\beta\in\mathcal{A}} \log \frac{\mathcal{P}_t^{\alpha\beta}}{\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}} \Bigg). \tag{22}$$

Using Definition 3.1 and reformulation steps in Section 4.1 from [22], (22) can be reformulated as:

$$\min_{\rho,\mathcal{P},\lambda,s} \mathbb{E}_\rho \sum_{t\in\mathcal{T}} \Bigg\{ \sum_{\alpha\in\mathcal{A}} -U_{t+1}^{\alpha} + \gamma \sum_{\beta\in\mathcal{A}} \Bigg(\sum_{\alpha\in\mathcal{A}} \log \mathcal{P}_t^{\alpha\beta} + \lambda\psi + \frac{1}{N}\sum_{y\in N} s_y \Bigg) \Bigg\} \tag{23a}$$

s.t.

$$\sup_{\substack{\overline{\mathcal{P}}^{\alpha\beta,\min} \leq \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} \leq \overline{\mathcal{P}}^{\alpha\beta,\max} \\ \sum_{\alpha\in\mathcal{A}} \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} = 1}} \sum_{\alpha\in\mathcal{A}} \{-\log \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} - \lambda |\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} - \overline{\mathcal{P}}_y^{\alpha\beta}|\} \leq s_y, \forall \beta \in \mathcal{A}, y \in N \tag{23b}$$

$$\text{Eq. (1b)} - \text{(1c)}, \tag{23c}$$

where s_y is an auxiliary variable and range $[\overline{\mathcal{P}}^{\alpha\beta,\min}, \overline{\mathcal{P}}^{\alpha\beta,\max}]$ defines the support for $\overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$, where parameters $\overline{\mathcal{P}}^{\alpha\beta,\min}$ and $\overline{\mathcal{P}}^{\alpha\beta,\max}$ are drawn from observations $\{\overline{\mathcal{P}}_y^{\alpha\beta}\}_{y\in N}$. Similarly to the relationship between (18) and (19), (23) represents an upper bound of (22) because it also maximizes over a convex function ($\sup - \log \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta}$). Since (23b) is convex, the supremum of (23b) can be obtained by an exhaustive search over extreme points. The extreme points are generated by the intersection of hyper-boxes, representing the range $[\overline{\mathcal{P}}^{\alpha\beta,\min}, \overline{\mathcal{P}}^{\alpha\beta,\max}]$, and a hyper-plane, ensuring that the probability of moving from present state β to all possible next states α is equal to one ($\sum_{\alpha\in\mathcal{A}} \overline{\boldsymbol{\mathcal{P}}}^{\alpha\beta} = 1$). This allows us to solve (23) using off-the-shelf solvers.

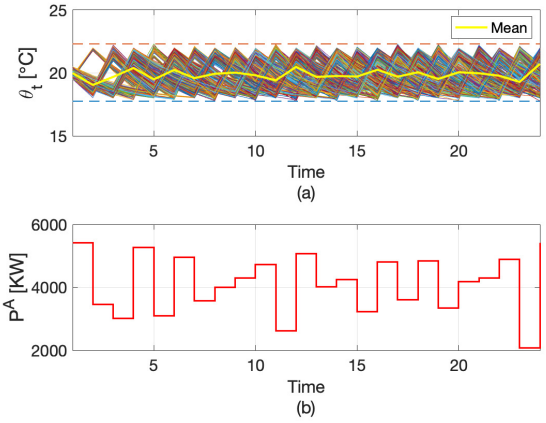

Figure 1. (a) Temperature evolution of the ensemble with 1000 TCLs and (b) their aggregated power consumption.

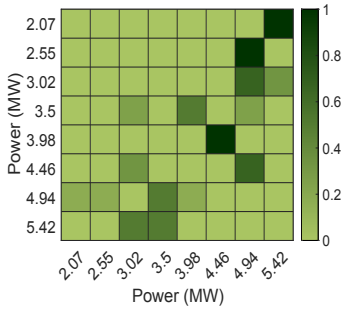

Figure 2. Default transition probability matrix ($\overline{\mathcal{P}}^{\alpha\beta}$) with 8 states constructed from the power profile in Fig. 1(b), where the color density indicates the probability value in the sidebar.

V. Case Study

The case study is carried out for a TCL ensemble with 1,000 residential air conditioner units. The discrete-time model for an individual residential air conditioner is based on [5], [33], [34] and given as:

$$\theta_{t+1} = \varrho\theta_t + (1-\varrho)(\theta^a - \aleph RPu_t) + \kappa_t, \tag{24}$$

where $\varrho = \exp(-h/RC)$, θ_t represents the indoor temperature of the room, θ^a is the ambient temperature, R is the thermal resistance, C is the thermal capacitance, P is the electrical power consumption, $u_t \in \{0,1\}$ determines whether the device is on or off, and $\aleph$ is the thermal efficiency. Parameter κ_t represents noise, which is ignored in the construction of the MP, and instead is accounted for by randomizing the default transition probabilities and solve it using different methods as given in Table I. Fig. 1 displays simulated temperature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2020.2992268, IEEE Transactions on Power Systems

Table II. COST PERFORMANCE OF ANALYTICAL CONTROL POLICIES.

Solution Cost (Objective function), $					
γ($)	η=0.00	η=0.25	η=0.50	η=0.75	η=1.00
0.05	2787.04	2786.63	2786.22	2785.81	2785.40
0.10	2805.10	2804.43	2803.76	2803.09	2802.42
1.00	2884.72	2879.16	2873.59	2867.99	2862.38

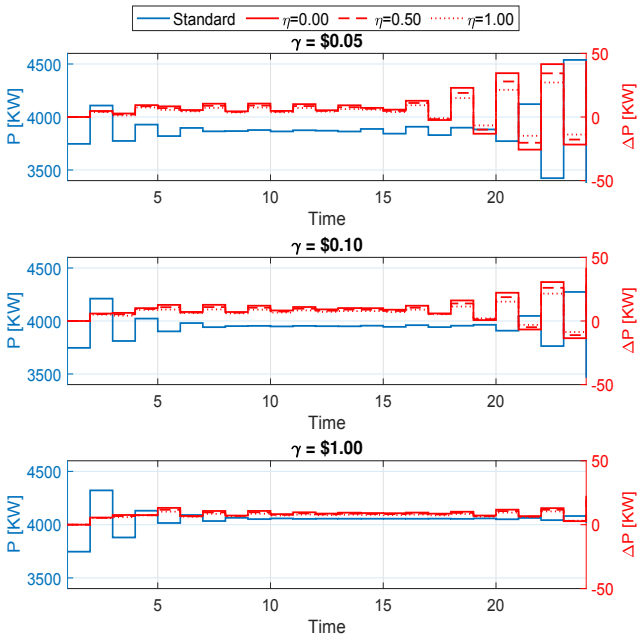

Figure 3. Optimal power dispatch under the standard MDP in (1) (blue) and the difference (denoted as ΔP) in the power consumption under analytical stochastic and distributionally robust control policies for different values of cost penalty γ. The stochastic and distributionally robust policies are computed using the hybrid model in (15) with $\eta = 1$ and $\eta = 0$, respectively.

trajectories and the resulting aggregated power consumption. The aggregated power consumption is discretized into 8 states with uniform power intervals and the associated probability transitions are shown in Fig. 2. These transitions are defined as the default transition probabilities ($\overline{\mathcal{P}}$) in our models. Next, we generate 1,000 random samples representing the set of observations by varying default transition probabilities within 15% of their nominal values in Fig. 2, while ensuring that the sum of probabilities remains equal to one. Then this set is used to estimate the empirical mean ($\overline{\mathcal{P}}^{\alpha\beta}$) and variance ($\sigma^2$) values. All simulations are performed using the Julia JuMP [35] package on an Intel Core i5 2.3 GHz processor with 8 GB of RAM and the Ipopt solver.

A. Analytical Control Policies

This section studies the performance and solution quality attained with the analytical control policies derived in Theorems 2–4. We implement the hybrid model and use it to obtain the stochastic and distributionally robust solutions by setting $\eta = 1.00$ and $\eta = 0.00$, respectively. For the mean and variance bounds in (10) and (11), we set the values of parameters $\xi = 0.001$ and $\varsigma = 0.1$. Table II summarizes the cost performance of all control policies for different values of η and γ and Fig. 3 itemizes the TCL ensemble power

Table III. COST PERFORMANCE OF ANALYTICAL CONTROL POLICIES IN THE DISTRIBUTIONALLY ROBUST CASE ($\eta = 0$).

Objective function, $				
γ($)	Parameter ς	Parameter ξ 0.1	0.01	0.001
0.05	0.1	**2785.34***	2786.25 (0.035% ↑)	2787.04 (0.061% ↑)
	0.10	2785.88 (0.019% ↑)	2786.88 (0.055% ↑)	2787.75 (0.086% ↑)
	0.001	2786.35 (0.036% ↑)	2787.42 (0.074% ↑)	2788.36 (0.108% ↑)
0.10	0.1	**2802.34***	2803.81 (0.052% ↑)	2805.10 (0.098% ↑)
	0.01	2803.16 (0.029% ↑)	2804.77 (0.086% ↑)	2806.17 (0.136% ↑)
	0.001	2803.87 (0.054% ↑)	2805.59 (0.115% ↑)	2807.09 (0.169% ↑)
1.00	0.1	**2861.87***	2874.05 (0.425% ↑)	2884.72 (0.798% ↑)
	0.01	2868.21 (0.221% ↑)	2881.42 (0.683% ↑)	2892.97 (1.086% ↑)
	0.001	2873.63 (0.410% ↑)	2887.71 (0.902% ↑)	2900.01 (1.332% ↑)

* Bold numbers are reference values.

dispatch[3] for selected values of η. As expected, the solution cost decreases as the value of parameter η increases, i.e. distributional robustness and the ability to accommodate high-fidelity assumptions on the underlying uncertainty come at a modest increase in the operating cost. However, the cost increases also depend on the value of chosen cost penalty γ. As γ increases, so does the cost difference between the stochastic and distributionally robust solutions. In terms of the power dispatch displayed in Fig. 3, internalizing the uncertainty on transition probabilities tends to increase the flexibility of the TCL ensemble[4] relative to the flexibility that can be extracted from the TCL ensemble relative to the standard MDP solution. In turn, the amount of this extra flexibility (ΔP in Fig. 3) depends on the time period and on the value of cost penalty γ. The greater this cost penalty, the less flexibility can be extracted from the TCL ensemble. We further evaluate the cost performance of the analytical control policies in the distributionally robust case ($\eta = 0$) for different mean and variance bounds by varying parameters ξ and ς in Table III and Fig. 4. It is observed that the solution cost increases with a decrease in values of ξ and ς, and the magnitude of this increase is greater for greater values of cost penalty γ. This is expected because decreasing the values of ξ and ς expands the confidence bounds around the mean and variance, which increases the robustness of the solution and immunizes it against a more extreme worst-case distribution.

Notably, the computational time for the analytical control policies in Theorems 2–4 is less than 0.013 seconds in all numerical experiments discussed above.

B. Numerical Control Policies

This section compares the cost and dispatch performance of distributionally robust solutions obtained using numerical

[3]Here and in the following discussions, the power dispatch is recovered from the MDP solution as $p_t = \sum_{\beta \in \mathcal{A}} p^{\beta,rated} \rho_t^{\beta}, \forall t \in \mathcal{T}$, where $p^{\beta,rated}$ is the rated power at each state and ρ_t^{β} is the MDP solution.

[4]In this case study, the term flexibility refers to the difference between the default power consumption and the power consumption with one of the proposed MDP solutions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2020.2992268, IEEE Transactions on Power Systems

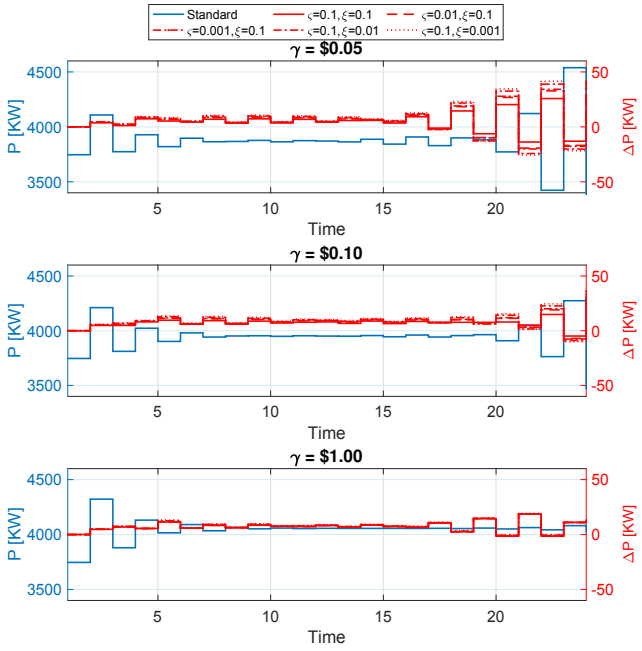

Figure 4. Optimal power dispatch under the standard MDP in (1) (blue) and the difference (denoted as ΔP) in the power consumption under the hybrid model ($\eta = 0$) in (15) for different values of cost penalty γ.

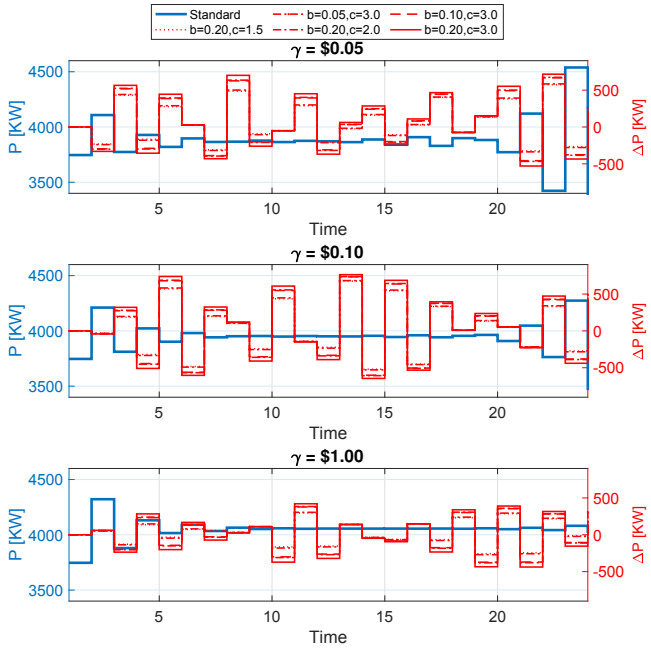

Figure 5. Optimal power dispatch under the standard MDP in (1) (blue) and the difference (denoted as ΔP) in the power consumption under the moment-based distributionally robust MDP in (20) (red) for different values of cost penalty γ.

Table IV. COST PERFORMANCE OF THE MOMENT-BASED MDP.

Objective function, $				
γ($)	Parameter c	Parameter b 0.05	0.10	0.20
0.05	1.5	**2766.81***	2840.93 (2.67% ↑)	2843.35 (2.76% ↑)
	2.0	2805.83 (1.41% ↑)	2883.15 (4.20% ↑)	2886.11 (4.31% ↑)
	3.0	2812.71 (1.65% ↑)	2891.76 (4.51% ↑)	2895.53 (4.65% ↑)
0.10	1.5	**2976.63***	3074.05 (3.27% ↑)	3077.51 (3.38% ↑)
	2.0	3021.09 (1.49% ↑)	3121.05 (4.85% ↑)	3126.17 (5.02% ↑)
	3.0	3032.11 (1.86% ↑)	3133.77 (5.27% ↑)	3139.94 (5.48% ↑)
1.00	1.5	**3179.43***	3295.54 (3.65% ↑)	3303.64 (3.90% ↑)
	2.0	3262.48 (2.61% ↑)	3382.84 (6.39% ↑)	3391.76 (6.67% ↑)
	3.0	3279.11 (3.13% ↑)	3401.34 (6.97% ↑)	3411.03 (7.28% ↑)

* Bold numbers are reference values.

Table V. COST PERFORMANCE OF THE WASSERSTEIN-BASED MDP.

Objective function, $			
γ($)	ψ=0.5	ψ=1.0	ψ=2.0
0.05	**2808.13***	2818.60 (0.37% ↑)	2836.59 (1.01% ↑)
0.10	**2814.84***	2826.33 (0.40% ↑)	2845.51 (1.08% ↑)
1.00	**2852.42***	2871.03 (0.65% ↑)	2902.16 (1.74% ↑)

* Bold numbers are reference values.

control polices described in Section IV. Tables IV and V present the solution cost for different values of parameter γ and Figures 5 and 6 compare the power dispatch of the TCL ensembles under moment- and Wasserstein-based ambiguity sets relative to the standard MDP formulation for different values of parameters b and c in (17) and ψ in (21). Naturally, the solution cost increases for greater values of cost penalty γ. Under both the moment- and Wasserstein-based ambiguity sets, the solution cost increases relative to the standard MDP and analytical control policies in Table II. These operating cost increases are expected, because using the ambiguous uncertainty sets makes it possible to better accommodate empirical observations, i.e. without assuming normally distributed errors on transition probabilities. In terms of the power dispatch, the moment-based approach leads to more volatile dispatch decisions for all values of cost penalty γ than in the Wasserstein-based case. Relative to the standard case, both the moment- and Wasserstein-based cases tend to increase the overall power flexibility (ΔP in Fig. 5 and 6) extracted from the TCL ensemble over 24 hours. Similar to the analytical control policies in Section V.A, we analyze the effects of confidence parameters on the cost performance of moment-based and Wasserstein-based methods. For the moment-based method, as presented in Table IV, the solution cost increases as the confidence region around the first and second-order moments widens by changing the values of parameters b and c. Fig. 5 displays the effect of varying b and c on the power dispatch of the TCL ensembles, where more inter-temporal fluctuations are observed for greater values of parameters b and c. In addition, in the Wasserstein-based method, we observe an increase in solution costs, see Table V, and power dispatch fluctuations, see Fig. 6, as radius ψ around the nominal distribution increases.

The average computational times for the moment- and Wasserstein-based cases are 18.2 and 44.5 seconds, which is significantly greater than for the analytical control policies.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2020.2992268, IEEE Transactions on Power Systems

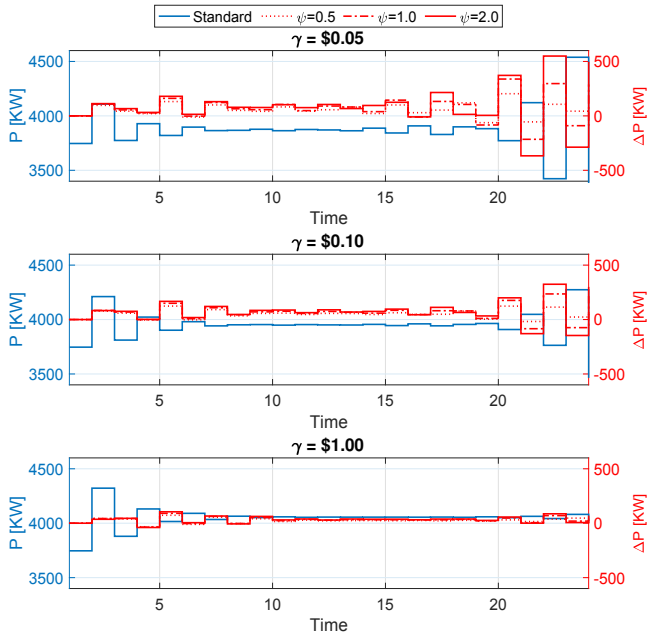

Figure 6. Optimal power dispatch under the standard MDP in (1) (blue) and the difference (denoted as ΔP) in the power consumption under the Wasserstein-based distributionally robust MDP in (23) (red) for different values of cost penalty γ.

VI. Conclusion

This paper describes analytical and numerical approaches to internalize the uncertainty dynamics of TCL ensembles in the Markov Decision Problem using stochastic and distributionally robust optimization. The stochastic and distributionally robust control policies are derived under mild assumptions on the underlying uncertainty and can be implemented in a computationally efficient manner. On the other hand, allowing for computationally demanding numerical control policies allows for better fitting empirical data, thus producing more accurate control policies and reducing data requirements for MDP problems. Our case study demonstrates that both the analytical and numerical control policies improve the accuracy of computing dispatch flexibility that can be extracted from the TCL ensemble relative to the standard MDP optimization, while minimizing the level of discomfort incurred to TCL users. Among different methods to accommodate the uncertainty in empirical measurements of TCL ensemble, we find that robust methods have more exogenous parameters that can be leveraged to intelligently trade-off solution cost and robustness. Although these exogenous parameters vary for the moment- and Wasserstein-based approaches, our numerical results demonstrate that they can be tuned in each case to achieve a comparable cost performance, thus allowing for distributionally robust decision-making in applications with different data availability.

Appendix A
Proofs of Theorems 1–4

We follow a similar procedure to prove all Theorems 1–4, where theorem-specific terms are denoted as $\mathcal{Z}$. The value of $\mathcal{Z}$ for each theorem is derived at the end of this appendix. For each theorem, given its respective MDP optimization, we can write the following Bellman equation $\forall t$ and $\forall \beta$:

$$\frac{1}{\gamma}\varphi_t^\beta = \frac{1}{\gamma}\min_{\mathcal{P}}\Big(-U_t^\beta + \mathbb{E}_{\mathcal{P}^{\alpha\beta}}\Big[\gamma\log\frac{\mathcal{P}_t^{\alpha\beta}}{\overline{\mathcal{P}}^{\alpha\beta}} + \mathcal{Z} + \varphi_{t+1}^\alpha\Big]\Big), \quad (25)$$

where φ_t^β is the value function at the present state β, φ_{t+1}^α is the value function from the next state α and $\mathcal{Z}$ represents a theorem-specific term for any possible transition probability uncertainty. Introducing the auxiliary (desirability) function $z_t^\beta = \exp(-\varphi_t^\beta/\gamma)$ in (25) leads to:

$$-\log(z_t^\beta) = \frac{1}{\gamma}\min_{\mathcal{P}}\Big(-U_t^\beta + \gamma\mathbb{E}_{\mathcal{P}^{\alpha\beta}}\Big[\log\frac{\mathcal{P}_t^{\alpha\beta}}{\overline{\mathcal{P}}^{\alpha\beta}} + \mathcal{Z} - \log(z_{t+1}^\alpha)\Big]\Big) =$$
$$\frac{1}{\gamma}\min_{\mathcal{P}}\Big(-U_t^\beta + \gamma\mathbb{E}_{\mathcal{P}^{\alpha\beta}}\Big[\log\frac{\mathcal{P}_t^{\alpha\beta}}{\overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^\alpha \exp(-\mathcal{Z})}\Big]\Big). \quad (26)$$

Next, the right-hand side of (26) is normalized using $\mathcal{G}^\beta(z) = \sum_\alpha \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^\alpha \exp(-\mathcal{Z})$, which results in:

$$-\log(z_t^\beta) = \frac{1}{\gamma}\min_{\mathcal{P}}\Big(-U_t^\beta + \gamma\mathbb{E}_{\mathcal{P}^{\alpha\beta}}\Big[\log$$
$$\frac{\mathcal{P}_t^{\alpha\beta}\mathcal{G}^\beta(z)}{\sum_\alpha \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^\alpha \exp(-\mathcal{Z})\mathcal{G}^\beta(z)}\Big]\Big) = \Big(\frac{-U_t^\beta}{\gamma} + \min_{\mathcal{P}}$$
$$\mathrm{KL}\Big[\mathcal{P}_t^{\alpha\beta}\Big\|\frac{\sum_\alpha \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^\alpha \exp(-\mathcal{Z})}{\mathcal{G}^\beta(z)}\Big] - \log\mathcal{G}^\beta(z)\Big), \quad (27)$$

where $\mathrm{KL}\big[\,\cdot\,\|\,\cdot\,\big]$ denotes the KL-divergence. The optimal policy is achieved when KL term in (27) is minimal, i.e. equal to zero. Since the zero value of the KL divergence is achieved when both distributions are identical, we obtain the condition for the optimal policy as:

$$\mathcal{P}_t^{\alpha\beta} = \frac{\overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^\alpha \exp(-\mathcal{Z})}{\mathcal{G}^\beta(z)} \quad (28)$$

Using the optimal policy in (28) and recalling that $\mathcal{G}^\beta(z) = \sum_\alpha \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^\alpha \exp(-\mathcal{Z})$, the Bellman equation in (27) can be recast as:

$$-\log(z_t^\beta) = \{-U_t^\beta/\gamma - \log\mathcal{G}^\beta(z)\} \quad (29)$$
$$\log(z_t^\beta) = \Big\{U_t^\beta/\gamma + \log\Big[\sum_\alpha \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^\alpha \exp(-\mathcal{Z})\Big]\Big\} \quad (30)$$

Exponentiating (30) leads to:

$$z_t^\beta = \exp\Big(U_t^\beta/\gamma\Big)\sum_\alpha \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^\alpha \exp(-\mathcal{Z}). \quad (31)$$

Since the value of $\mathcal{Z}$ varies for Theorems 1-4, we derive theorem-specific results for each case below.

A. Standard Formulation in Theorem 1

The standard model ignores the uncertainty of transition probabilities, which leads to:

$$\mathcal{Z}^S := \mathcal{Z} = 0. \quad (32)$$

Accordingly, using (32) returns the following optimal policy:

$$\mathcal{P}_t^{\alpha\beta} = \frac{\overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^{\alpha}}{\sum_{\alpha} \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^{\alpha}}. \tag{33}$$

B. Stochastic Formulation in Theorem 2

The value of $\mathcal{Z}$ for the stochastic model follows from (7a) as:

$$\mathcal{Z}^{E} := \mathcal{Z} = \frac{(\gamma\sigma^2)}{2(\overline{\mathcal{P}}^{\alpha\beta})^2} \tag{34}$$

Accordingly, using (34) returns the following optimal policy:

$$\mathcal{P}_t^{\alpha\beta} = \frac{\overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^{\alpha} \exp\left(\frac{-\gamma\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\right)}{\sum_{\alpha} \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^{\alpha} \exp\left(\frac{-\gamma\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\right)}. \tag{35}$$

C. Distributionally Robust Formulation in Theorem 3

The value of $\mathcal{Z}$ for the distributionally robust formulation follows from (13a) as:

$$\mathcal{Z}^{WC} := \mathcal{Z} = \frac{(\gamma\overline{\hat{\zeta}})}{2(\underline{\Gamma})^2} \tag{36}$$

Accordingly, using (36) returns the following optimal policy:

$$\mathcal{P}_t^{\alpha\beta} = \frac{\underline{\Gamma} z_{t+1}^{\alpha} \exp\left(\frac{-\gamma\overline{\hat{\zeta}}}{2(\underline{\Gamma})^2}\right)}{\sum_{\alpha} \underline{\Gamma} z_{t+1}^{\alpha} \exp\left(\frac{-\gamma\overline{\hat{\zeta}}}{2(\underline{\Gamma})^2}\right)}. \tag{37}$$

where $\overline{\mathcal{P}}^{\alpha\beta}$ is replaced by its bound $\underline{\Gamma}$ from the set $\mathbb{D}$ to obtain the worst-case distribution.

D. Hybrid Model in Theorem 4

Using (35) and (37), the hybrid optimal policy follows as:

$$\begin{aligned}\mathcal{P}_t^{\alpha\beta} = (1-\eta)&\frac{\underline{\Gamma} z_{t+1}^{\alpha} \exp\left(\frac{-\gamma\overline{\hat{\zeta}}}{2(\underline{\Gamma})^2}\right)}{\sum_{\alpha} \underline{\Gamma} z_{t+1}^{\alpha} \exp\left(\frac{-\gamma\overline{\hat{\zeta}}}{2(\underline{\Gamma})^2}\right)} \\ &+ \eta\frac{\overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^{\alpha} \exp\left(\frac{-\gamma\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\right)}{\sum_{\alpha} \overline{\mathcal{P}}^{\alpha\beta} z_{t+1}^{\alpha} \exp\left(\frac{-\gamma\sigma^2}{2(\overline{\mathcal{P}}^{\alpha\beta})^2}\right)}\end{aligned} \tag{38}$$

where $0 \leq \eta \leq 1$.

REFERENCES

[1] The Brattle Group, "The National Potential for Load Flexibility: Vlaue and Market Potential through 2030," 2019. [Online]. Available: shorturl.at/beP06

[2] M. Chertkov and V. Chernyak, "Ensemble of thermostatically controlled loads: statistical physics approach," *Sc. Rep.*, vol. 7, no. 1, p. 8673, 2017.

[3] FERC, "2018 assessment of demand response and advanced metering," 2018. [Online]. Available: shorturl.at/gwzBC

[4] V. Lakshmanan et al., "Impact of thermostatically controlled loads' demand response activation on aggregated power: A field experiment," *Energy*, vol. 94, pp. 705–714, 2016.

[5] S. Koch, J. L. Mathieu, and D. S. Callaway, "Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services," in *Proc. Power Systems Computation Conf*, 2011.

[6] A. Bušić and S. Meyn, "Ordinary differential equation methods for markov decision processes and application to kullback–leibler control cost," *SIAM J. Cont. and Opt.*, vol. 56, no. 1, pp. 343–366, 2018.

[7] M. Chertkov, V. Y. Chernyak, and D. Deka, "Ensemble control of cycling energy loads: Markov decision approach," in *Energy Markets and Responsive Grids: Modeling, Control, and Optimization*. vol 162. Springer New York, 2018, pp. 363–382.

[8] M. Chertkov, D. Deka, and Y. Dvorkin, "Optimal ensemble control of loads in distribution grids with network constraints," in *Power Systems Computation Conference (PSCC), Dublin, Ireland*, June 2018, pp. 1–7.

[9] E. Benenati, M. Colombino, and E. Dall'Anese, "A tractable formulation for multi-period linearized optimal power flow in presence of thermostatically controlled loads," *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 4189–4194, 2019.

[10] A. Hassan et al., "Optimal load ensemble control in chance-constrained optimal power flow," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5186–5195, Sep. 2019.

[11] K. Turitsyn, S. Backhaus, M. Ananyev, and M. Chertkov, "Smart finite state devices: A modeling framework for demand response technologies," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, Dec 2011, pp. 7–14.

[12] S. P. Meyn, P. Barooah, A. Bušić, Y. Chen, and J. Ehren, "Ancillary service to the grid using intelligent deferrable loads," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2847–2862, Nov 2015.

[13] S. Meyn, P. Barooah, A. Bušić, and J. Ehren, "Ancillary service to the grid from deferrable loads: The case for intelligent pool pumps in florida," in *52nd IEEE Conference on Decision and Control*, 2013, pp. 6946–6953.

[14] M. Vrakopoulou, B. Li, and J. L. Mathieu, "Chance constrained reserve scheduling using uncertain controllable loads part i: Formulation and scenario-based analysis," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1608–1617, March 2019.

[15] D. S. Callaway, "Tapping the energy storage potential in electric loads to deliver load following and regulation," *Energy Conversion and Management*, vol. 50, no. 5, pp. 1389–1400, 2009.

[16] S. Mannor et al., "Bias and variance approximation in value function estimates," *Management Science*, vol. 53, no. 2, pp. 308–322, 2007.

[17] E. Delage and S. Mannor, "Percentile optimization for markov decision processes with parameter uncertainty," *Op. Res.*, vol. 58, 2010.

[18] A. Nilim and L. El Ghaoui, "Robustness in markov decision problems with uncertain transition matrices," in *Adv. in NIPS*, 2004, pp. 839–846.

[19] H. Xu and S. Mannor, "Distributionally robust markov decision processes," in *Advances in NIPS*, 2010, pp. 2505–2513.

[20] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust markov decision processes," *Math. of Oper. Res.*, vol. 38, no. 1, pp. 153–183, 2013.

[21] E. Todorov, "Linearly-solvable markov decision problems," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 1369–1376.

[22] P. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Math. Program.*, vol. 171, pp. 115–166, Sep. 2018.

[23] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," 2016.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.

[25] T. Warren Liao, "Clustering of time series data-a survey," *Pattern Recogn.*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005.

[26] D. Métivier and M. Chertkov, "Mean-field control for efficient mixing of energy loads," *Phys. Rev. E*, vol. 101, p. 022115, Feb 2020. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.101.022115

[27] Y.W Teh et al., "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, 2006, pp. 1353–1360.

[28] C. Walck, *Hand-book on statistical distributions for experimentalists*, 1996.

[29] P. M. Young and M. A. Dahleh, "Infinite-dimensional convex optimization in optimal and robust control theory," *IEEE Transactions on Automatic Control*, vol. 42, no. 10, pp. 1370–1381, Oct 1997.

[30] R. Rockafellar, *Conjugate Duality and Optimization*. Society for Industrial and Applied Mathematics, 1974.

[31] S. K. Mitter, "Convex optimization in infinite dimensional spaces," in *Recent Advances in Learning and Control*, V. D. Blondel, S. P. Boyd, and H. Kimura, Eds. London: Springer London, 2008, pp. 161–179.

[32] A. Shapiro, *On Duality Theory of Conic Linear Problems*. Boston, MA: Springer US, 2001, pp. 135–165.

[33] R. E. Mortensen and K. P. Haggerty, "A stochastic computer model for heating and cooling loads," *IEEE Tran. Power Syst.*, vol. 3, no. 3, pp. 1213–1219, Aug 1988.
[34] S. Acharya et al., "Coordinated frequency control strategy for an islanded microgrid with demand side management capability," *IEEE Tran. Energy Conv.*, vol. 33, no. 2, pp. 639–651, June 2018.
[35] I. Dunning, J. Huchette, and M. Lubin, "Jump: A modeling language for mathematical optimization," *SIAM Rev.*, vol. 59, pp. 295–320, 2017.