A General Framework for Learning-Based Distributionally Robust MPC of Markov Jump Systems

Mathijs Schuurmans and Panagiotis Patrinos

Abstract—In this article, we present a data-driven learning model predictive control (MPC) scheme for chance-constrained Markov jump systems with unknown switching probabilities. Using samples of the underlying Markov chain, ambiguity sets of transition probabilities are estimated, which include the true conditional probability distributions with high probability. These sets are updated online and used to formulate a time-varying, risk-averse optimal control problem. We prove recursive feasibility of the resulting MPC scheme and show that the original chance constraints remain satisfied at every time step. Furthermore, we show that under sufficient decrease of the confidence levels, the resulting MPC scheme renders the closed-loop system mean-square stable with respect to the true-but-unknown distributions, while remaining less conservative than a fully robust approach. Finally, we show that the data-driven value function of the learning MPC converges from above to its nominal counterpart as the sample size grows to infinity. We illustrate our approach on a numerical example.

Index Terms—Predictive control, robustness, statistical learning.

I. INTRODUCTION

A. Background, Motivation and Related Work

Due to the ubiquitous nature of stochastic uncertainty in processes arising in virtually all branches of science and engineering, control of dynamical systems perturbed by stochastic processes is a long-standing topic of research. Model predictive control (MPC)—stochastic MPC in particular—has been a popular and successful tool in this endeavor, due to its ability to naturally include probabilistic information directly into the control design via the cost, the dynamics, and the constraints [2], [3], [4]. In classical stochastic MPC, however, it is typically assumed that the distribution of the underlying stochastic process is known, although in practice, this is usually not the case. If the disturbance takes values on a bounded set, the absence of full distributional knowledge can be taken into account by designing the controller under the worst-case realization of the stochastic disturbance. This approach is referred to as robust MPC [2], [4].

An obvious drawback of robust approaches is that the complete disregard of the probabilistic nature of the disturbance can be rather crude, resulting in a tendency for overly conservative decisions. As an alternative approach, one may simply compute an empirical estimate of the disturbance distribution and replace the true value by this estimate in the optimal control problem. Although this is a reasonable approach given a sufficient amount of data, for more moderate sample sizes, there may be a significant misestimation of the underlying distributions—often referred to as ambiguity. It is well known that this is likely to cause degradation of the resulting performance when evaluated on new samples from the true distribution. This phenomenon is known as the optimizer’s curse [5]. To account for this ambiguity, one could, instead of a point estimate, construct a set of all distributions (an ambiguity set) that are in some specific sense consistent with the data. By accounting for the worst-case distribution within this set, the decision maker is protected against the limitations of the finite sample size.

This approach, known as distributionally robust (DR) optimization [6], addresses the drawbacks of the abovementioned approaches by utilizing available data, but only to the extent that it is statistically meaningful. As more data are gathered online and ambiguity sets get updated accordingly, it is expected that these sets will shrink, so that the optimal decisions gradually become less conservative. This, among other desirable properties, has caused an increasing popularity of DR methods in recent years, initially mostly in stochastic programming and operations research communities [5], [7], [8], [9], [10] and more recently in (optimal) control [11], [12], [13], [14], [15], [16] as well. See also [17] for a comprehensive review. Much of the earlier work focuses on the study of particular classes of ambiguity sets, each modeling certain structural assumptions on the underlying distribution. Although most of our analysis does not require a particular family of ambiguity sets, we will, for concreteness, put particular emphasis on ambiguity sets that are written as a divergence ball around an empirical estimate, as this family of sets is a natural choice in the setting at hand. This is described...
in Section III, where a table containing several choices for the divergence is provided.

As the focus of research in data-driven and learning-based control is gradually shifting toward real-life, safety-critical applications, there has been an increasing concern for safety guarantees of data-driven methods, which are valid in a finite data regime. This has led to a variety of different approaches besides DR methodologies, each valid under different assumptions on the data-generating process and the controlled systems. For instance, this has led to data-driven variants of tube-based MPC [18], [19], Gaussian-process based estimation with reachability-based safe set constraints [20], data-enabled predictive control [21] combining Willems’ fundamental lemma with MPC for linear systems, or techniques based on Koopman operators [22]. We refer to [23] for a recent survey.

In this work, we allow for general (possibly nonlinear) dynamics under stochastic disturbances with unknown distribution, and subject to chance constraints. However, we restrict our attention to finitely supported stochastic disturbances. One of the advantages of this construction is that the predicted evolution of the system can be represented on a scenario tree, which allows us to explicitly (and without approximation) optimize over closed-loop control policies, rather than open-loop sequences. This property helps combat excessive conservatism due to accumulation of uncertainty over the prediction horizon [24], [25], [26]. Motivated by similar considerations, Leider et al. [27] and Bonzanini et al. [28] utilized scenario trees to approximate the realizations of continuous disturbances. Bonzanini et al. [28] then considered safety separately by projecting the computed control action onto a set of control actions that keep the state within safe robust control invariant (RCI) set, similarly to [20]. This projection requires the additional solution of a mixed-integer quadratic program (MIQP), whenever the used RCI set is polyhedral. In our setting, however, we consider the switching behavior inherent to the system, allowing us to provide safety guarantees directly through the application of MPC theory on the joint controller–learner system.

We will in particular assume that the underlying disturbance process is a Markov chain, leading to a system class commonly referred to as Markov jump systems. Control of this class of systems has been widely studied and has been used to model systems stemming from a wide range of applications [25], [29], [30]. In the known distribution case, the stability analysis of nonlinear stochastic MPC for this system class has been performed from a worst-case perspective [31], in mean-square sense [30] and in the more general risk-square sense [32], [33]. We emphasize here the distinction between risk averse and DR approaches, where the former optimizes a given coherent risk measure with respect to the true distribution, whereas the latter constructs a data-driven ambiguity set with respect to which the stochastic cost is robustified. By the dual risk representation [34, Th. 6.4], every ambiguity set induces some coherent risk measure and vice versa, leading both approaches to solve the same class of optimization problems. However, the statistical interpretation and, thus, the corresponding guarantees differ significantly.

Indeed, by the mentioned equivalence, the notion of risk-square stability in [32] guarantees mean-square stability (MSS) with respect to all the distributions within the “ambiguity set” induced by the used risk measure. In practice, however, this is insufficient to guarantee MSS with respect to the true-but-unknown distribution, as it is impossible to construct a non-trivial ambiguity set that contains the true distribution with certainty. However, we will show that by the careful design of a data-driven sequence of ambiguity sets—which only contain the true distributions with high probability—this concept can be extended to show MSS, as well as recursive constraint satisfaction with respect to the true distribution, under some additional assumptions.

Other data-driven methods have been proposed to design controllers for unknown transition probabilities [35], [36]. However, these works are restricted to a simpler, unconstrained setting involving only linear state-feedback policies. Furthermore, related risk-averse and DR techniques have been proposed for Markov decision processes (MDPs) [37], [38], [39], [40], although these consider discrete states and actions, allowing one to solve directly the Bellman equation over all admissible policies. Unfortunately, these techniques become intractable in the present setting involving continuous states and actions.

We finally study the convergence of the optimal value function of our learning controller to the nominal counterpart. This property, known as asymptotic consistency, has recently been studied in the stochastic optimization literature for (static) DR optimization problems under Wasserstein ambiguity [5], [41]. A common assumption in this line of work is Lipschitz continuity of the cost/constraint functions with respect to the random variable. This assumption is not suitable for our purposes, since we consider discrete random variables $w \in W$ for which a suitable norm may not exist. Instead, we will in some cases need to resort to a uniform boundedness assumption, which serves a similar purpose. In the nonconvex case, Cherukuri and Hota [41] based their analysis on [42], in which the ambiguity sets are not assumed to be random. An additional assumption is added that the constraint boundary has probability zero, such that almost everywhere, the constraint is continuous. This assumption helps in dealing with the discontinuity of the step function at 0, which is inherent to chance constraints. Alternatively, the chance constraints can be replaced by risk constraints involving the average value-at-risk [43], which also circumvents this issue. Besides the mentioned differences in assumptions, additional care is required to handle the multistage nature of the stochastic optimization problems considered here. Specifically, both the optimal cost and the feasible set are defined recursively through the Bellman operator (see Section V), causing more complex characterizations of the optimal value function as well as reduced freedom in selecting the problem parameters to ensure its required properties as compared with a static two-stage stochastic program.

B. Contributions

Summarizing the previous discussion, we highlight the following contributions of our work.

1) We present a general online learning DR-MPC framework for Markov switching systems with unknown transition probabilities. The resulting closed-loop system satisfies the (chance) constraints of the original stochastic problem and allows for online improvement of performance based on observed data. Thus, we extend the recently developed framework of risk-averse MPC [32], [33], [44] to an online learning setting, in which the involved risk measures are selected and calibrated automatically based on their dual (DR) interpretation. To this end, we formalize the procedure for estimating and updating the corresponding ambiguity sets as a dynamical system, which we refer to as the learning system. We present conditions on this learning system to ensure its convergence and
to obtain meaningful statistical guarantees on the resulting controllers with respect to the unknown underlying distributions.

2) We provide sufficient conditions for recursive feasibility and MSS of the DR-MPC law, with respect to the true-but-unknown distribution. To this end, we state the problem in terms of an augmented state vector, including the state of the previously mentioned learning system. The dynamics of this so-called learner state can be easily expressed for common choices for the ambiguity set. This idea, which is closely related to that of information states [45, Ch. 5], allows us to formulate the, otherwise, time-varying optimal control problem as a dynamic programming recursion, facilitating stability analysis of the original control system and the learning system jointly.

3) We provide sufficient conditions under which the value of the DR problem converges from above to that of the nominal optimal control problem, extending existing results in stochastic optimization to the constrained, multistage, dynamical setting.

C. Notation

Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{N}_{\geq 0} := \mathbb{N} \setminus \{0\}$. For two naturals $a, b \in \mathbb{N}$ with $a \leq b$, we denote $[a,b] := \{n \in \mathbb{N} \mid a \leq n \leq b\}$, and similarly, we introduce the shorthand $w_{[a,b]} := (w_t)_{t=a}^{b}$ to denote a sequence of variables indexed from $a$ to $b$. We denote the extended real line by $\mathbb{R} := \mathbb{R} \cup \{\pm \infty\}$ and the set of nonnegative (extended) real numbers by $\mathbb{R}_{+}$ (and $\mathbb{R}_{+}^{\infty}$). The cardinality of a (finite) set $W$ is denoted by $|W|$. We write $f : X \rightrightarrows Y$ to denote that $f$ is a set-valued mapping from $X$ to $Y$. A function is lower semicontinuous (lsc) if its epigraph is closed. Given a matrix $P \in \mathbb{R}^{n \times m}$, we denote its $(i,j)^{th}$ element by $P_{ij}$ and its $i^{th}$ row as $P_i \in \mathbb{R}^{m}$. The $i^{th}$ element of a vector $x$ is denoted $x_i$, $\text{vec}(M)$ denotes the vertical concatenation of the columns of a matrix $M$. We denote the vector in $\mathbb{R}^k$ with all elements one as $1_k := (1)_{i=1}^{k}$ and the probability simplex of dimension $k$ as $\Delta_k := \{ p \in \mathbb{R}_+^k \mid p^T 1_k = 1 \}$. We define the function $1_{x=y} = 1$ if $x = y \in \mathbb{R}$, otherwise. The indicator function $\delta_X : \mathbb{R}^n \rightarrow \mathbb{R}$ of a set $X \subseteq \mathbb{R}^n$ is defined by $\delta_X(x) = 0$ if $x \in X$ and $\infty$, otherwise. The set of a function $V : \mathbb{R}^n \rightarrow \mathbb{R}$ is denoted $\text{lev}_{\leq \alpha} V := \{ x \in \mathbb{R}^n \mid V(x) \leq \alpha \}$. The interior of a set $X$ is denoted $\text{int} X$. We denote the positive part of a quantity $x$ as $[x]_+ := \max\{0,x\}$, where max is taken elementwise. We say that a function $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ belongs to the class of $K_{\infty}$ functions if it is continuous, strictly increasing, unbounded, and zero at zero [4]. Finally, given a nonempty, proper cone $K$, the generalized inequality $a \leq_K b$ is equivalent to $b - a \in K$. $K^\circ := \{ y \mid \langle x,y \rangle \geq 0 \ \forall x \in K \}$ denotes the dual cone of $K$.

II. PROBLEM STATEMENT AND STRUCTURAL ASSUMPTIONS

Let $w := (w_t)_{t \in \mathbb{N}}$ denote a discrete-time, time-homogeneous Markov chain defined on some probability space $\Omega, (F, P)$ and taking values on $W := \mathbb{N}_{[1,d]}$. The transition kernel governing the Markov chain is denoted by $P = (P_{ij})_{i,j \in W}$, where $P_{ij} = P[w_t = j \mid w_{t-1} = i]$. We refer to $w_t$ as the mode of the chain at time $t$. For simplicity, we will assume that the initial mode is known to be $i$, so $p_0 = (1_{w=i})_{w \in W}$. Therefore, the Markov chain is fully characterized by its transition kernel. Finally, we will assume that the Markov chain is ergodic.

Assumption II.1 (Ergodicity): The Markov chain $(w_k)_{k \in \mathbb{N}}$ is ergodic, i.e., there exists a value $k \in \mathbb{N}_{>0}$, such that $P^k > 0$ elementwise.

This assumption, stating that every mode is reachable from any other mode in $k$ steps, ensures that every mode of the chain gets visited infinitely often [46, Ex. 8.7]. This will allow us to guarantee convergence of the proposed learning MPC scheme to its nominal counterpart. (See Section V.)

We will consider discrete-time systems with dynamics of the form

$$x_{t+1} = f(x_t, u_t, w_{t+1}) \quad (1)$$

where $x_t \in \mathbb{R}^{n_x}$ and $u_t \in \mathbb{R}^{n_u}$ are the state and control action at time $t$, respectively. We will assume that the state $x_t$ and mode $w_t$ are observable at time $t$. This is equivalent to the more common notation $x_{t+1} = f(x_t, u_t, w_t)$, assuming $w_{t-1}$ is observable. However, as we will consider $w_t$ to be a part of the system state at time $t$, the notation of (1) will be more convenient.

Since $w_t$ is drawn from a Markov chain, such systems are commonly referred to as Markov jump systems. Whenever $f(\cdot, \cdot, w)$ is a linear function, (1) describes a Markov jump linear system [29]. Since the state $x_t$ and mode $w_t$ are observable at time $t$, the distribution of $x_{t+1}$ depends solely on the conditional switching distribution $P_{w_t \mid w}$, for a given control action $u_t$.

For a given state-mode pair $(x, w)$, we will impose probabilistic constraints of the form

$$\text{AV}\@\mathbb{R}_+^{n_w} \{ g_t(x, u, w, v) \mid x, w \leq 0, \ i \in \mathbb{N}_{[1,n_g]} \} \quad (2)$$

where $v \sim P_{w}$ is randomly drawn from the Markov chain $w$ in mode $w$, $g_t : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^2 \rightarrow \mathbb{R}$ are constraint functions with corresponding constraint violation rates $\alpha_t$, and $\text{AV}\@\mathbb{R}$ denotes the (conditional) average value-at-risk. The conditional $\text{AV}\@\mathbb{R}$ (at level $\alpha \in [0,1]$ and with reference distribution $p \in \Delta_d$) of the random variable $\xi : \mathbb{W} \rightarrow \mathbb{R}$ is defined as

$$\text{AV}\@\mathbb{R}_+^{n_w} \{ \xi(w, v) \mid w \} = \begin{cases} \min_{t \in \mathbb{R}} +1/\alpha \ E_p \{ (\xi(w, v) - t)_{+} \mid w \}, & \alpha \neq 0 \\ \max_{w \in \mathbb{W}} \{ \xi(w, v) \}, & \alpha = 0 \end{cases} \quad (3)$$

and it has the useful property that if $p = P_{w}$, then the following implication holds tightly [34, Sec. 6.2.4]:

$$\text{AV}\@\mathbb{R}_+^{n_w} \{ \xi(w, v) \mid w \leq 0 \Rightarrow P[\xi(w, v) \leq 0 \mid w] \geq 1 - \alpha. \} \quad (4)$$

By exploiting the dual risk representation [34, Th. 6.5], the left-hand inequality in (4) can be formulated in terms of only linear constraints [44]. As such, it can be used as a tractable surrogate for chance constraints that would lead to nonconvex, nonsmooth constraints [43]. By appropriate choices of $\alpha_t$ and $g_t$, constraint (2) can be used to encode robust constraints ($0 < \alpha_t < 1$) on the state, the control action, or both. Note that it additionally covers chance constraints on the successor state $f(x, u, v)$ under input $u$, conditioned on the current values $x$ and $w$. To ease notation, we will without loss of generality assume that $n_g = 1$. To summarize, the set of feasible
control actions as a function of $x$ and $w$ can be written as

$$\mathcal{U}(x, w) := \{ u \in U : AV@ R^n_{ux} [g(x, u, w, v) \mid x, w] \leq 0 \}$$

(5)

where $U \subseteq \mathbb{R}^{n_u}$ is a nonempty, closed set.

Ideally, our goal is to synthesize—by means of a stochastic MPC scheme—a stabilizing control law $\kappa_N : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathcal{P}_\mathbb{R}^{n_x}$, such that for the closed-loop system $x_{t+1} = f(x_t, \kappa_N(x_t, w_t), w_{t+1})$, it holds almost surely that $\kappa_N(x_t, w_t) \in \mathcal{U}(x_t, w_t)$, for all $t \in \mathbb{N}$. Consider a sequence of $N$ control laws $\pi = (\pi_N)_{N=0}^{\infty}$, referred to as a policy of length $N$. Given a stage cost $\ell : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathcal{R}_+$ and a terminal cost $V_f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathcal{R}_+$ and corresponding terminal set $\Delta_\pi := \left\{ V_f(x, w) + \delta_\pi(x, w) \right\}$, we can assign to each such policy $\pi$, a cost

$$V_N(x, w) := E \left[ \sum_{k=0}^{N-1} \ell(x_k, u_k, x_{k+1}) + V_f(x_N, w_N) \right]$$

(6)

where $x_{k+1} = f(x_k, u_k, w_{k+1})$, and $u_k = \pi_k(x_k, w_k)$ and $(x_0, w_0) = (x, w)$, for $k \in \mathbb{N}_{[0,N-1]}$. This defines the following stochastic OCP.

**Definition II.2 (Stochastic OCP):** For a given state-mode pair $(x, w)$, the optimal cost of the stochastic OCP is

$$V_N(x, w) = \min_{\pi} V_N^\pi(x, w)$$

subject to

$$x_0 = x, w_0 = w, \pi = (\pi_N)_{N=0}^{\infty} = (\pi_k)_{k=0}^{N-1}$$

(7a)

$$x_{k+1} = f(x_k, \pi_k(x_k, w_k), w_{k+1})$$

(7b)

$$\pi_k(x_k, w_k) \in \mathcal{U}(x_k, w_k) \quad \forall k \in \mathbb{N}_{[0,N-1]}.$$  

(7c)

We denote by $\Pi_N(x, w)$ the corresponding set of minimizers.

To ensure existence of a solution to (7) (and its DR counterpart, defined in Section IV), we will impose the following (standard) regularity conditions [41, 32].

**Assumption II.3 (Problem regularity):** The following are satisfied for all $u, v \in \mathcal{W}$:

i) functions $\ell(\cdot, \cdot, \cdot) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}_+$, $V_f(\cdot, \cdot, \cdot) : \mathbb{R}^{n_x} \to \mathbb{R}_+$, and $g(\cdot, \cdot, \cdot, \cdot) : i \in \mathbb{N}_{[1,n_{ox}]}$ are continuous;

ii) $U$ and $\mathcal{X}_t$ are closed;

iii) $f(0, 0, w) = 0, \ell(0, 0, w) = 0, \ell(U(0, w), x) = 0, V_f(0, w) = 0$;

iv) one of the following is satisfied: 1) $U$ is compact;

2) $\ell(x, u, w) \geq c(\|u\|)$ with $c \in \mathcal{K}_{\infty}, \forall (x, u) \in \mathbb{R}^{n_x} \times U$. Let $(\pi^*_N(x, w))_{N=1}^{\infty} \subseteq \Pi_N(x, w)$, so that the stochastic MPC control law is given by $\kappa_N(x, w) = \pi^*_N(x, w)$. Sufficient conditions on the terminal cost $V_f$ and its effective domain $\text{dom} \mathcal{V}_t$ to ensure MSS of the closed-loop system have been studied for a similar problem set-up in [30], among others.

Both designing and computing such a stochastic MPC law requires knowledge of the probability distribution governing the state dynamics (1), or equivalently, of the transition kernel $P$. In the absence of this knowledge, these probabilities are to be estimated from a finitely sized dataset, and therefore subject to some level of ambiguity. Our goal is to devise an MPC scheme that uses the available data in a principled manner, while explicitly taking this ambiguity into account.

To this end, we will introduce the notion of a learner state, which is very similar in spirit to the concept of an information state, commonly used in control of partially observed MDPs [47], where—in contrast to our approach—it is typically adopted in a Bayesian setting. In both cases, however, it can be regarded as an internal state of the controller that stores all the information required to build (at least) one conditional distribution over the next state, given the observed data. We will make this more precise in the next section. Equipped with such a learning system, our aim is to find a data-driven approximation to the stochastic OCP defined by (7), which asymptotically attains the optimal cost while preserving stability and constraint satisfaction during closed-loop operation.

The rest of this article is organized as follows. Section III formalizes the assumed learning system and presents and several classes of ambiguity sets found in the literature that fit the framework. In Section IV, we use this learning system to construct a DR counterpart to the optimal control problem in terms of the ingredients introduced above. Section V contains a theoretical analysis of the proposed scheme. In Section VI, we illustrate the approach on some numerical examples. Finally, Section VII concludes this article.

III. DATA-DRIVEN AMBIGUITY SETS

A. Abstract Learning System

As mentioned in the previous section, we model the procedure that maps the observed data into a set of transition probabilities as a generic Markovian system, which we refer to as the learning system. We first state the required structure in a compact, abstract notation and later provide a concrete example, which will suffice in many practical cases.

**Assumption III.1 (Learning system):** Given a sequence $w_{0:t}$ sampled from the Markov chain $w$, we can compute the following:

1) a statistic $s_t : W^{t+1} \to S \subseteq \mathbb{R}^{n_s}$, with $S$ compact, accompanied by a vector of confidence parameters $\beta_t = (\beta_{t,i})_{i=1}^{n_s} \in \mathcal{I} := [0, 1]^{n_s}$, for which there exist some Markovian dynamics $\mathcal{L}$ and $C$ such that $s_{t+1} = \mathcal{L}(s_t, \beta_t, w_t, w_{t+1})$ and $\beta_{t+1} = C(\beta_t), t \in \mathbb{N}$;

2) an ambiguity set $A : S \times W \times [0, 1] \to \Delta_{\mathcal{L}} : (s, w, \beta) \mapsto A_{\beta}(s, w)$, mapping $s_t, w_t$, and the component $\beta_{t,i}$ to a convex subset of the $d$-dimensional probability simplex $\Delta_d$, such that for all $t \in \mathbb{N}, w \in W, i \in \mathbb{N}_{[1,n_{eta}]}$

$$\mathbb{P} [P_{w} \in A_{\beta_{t},i}(s_t, w)] \geq 1 - \beta_{t,i}.$$  

(8)

We will refer to $s_t$ and $\beta_t$ as the learner state and the confidence vector at time $t$, respectively.

**Remark III.2 ( Learner dynamics $\mathcal{L}$, $C$):** The existence of the dynamics $\mathcal{L}$ and $C$ implies that the system with the augmented state consisting of both the original system state-mode pair $(x_t, v_t)$ and the learner-confidence pair $(s_t, \beta_t)$ is Markovian. This assumption aids the theoretical analysis in Section V and is not restrictive in practice, as it essentially only requires that finite memory is needed for the method, which is the case for all implementable methods. For concreteness, typical examples for $\mathcal{L}$ and $C$, which are valid for many practical use cases, are presented in Examples III.8 and III.6, respectively.

**Remark III.3 (Confidence levels):** We consider a vector of confidence levels, rather than a single value. This is motivated by the fact that one would often wish to assign separate confidence levels to ambiguity sets corresponding to the cost function, and to those corresponding to the $n_s$ chance constraints (See Definition IV.3). Accordingly, we will assume that $n_s = n_s + 1$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
In order to ensure reasonable behavior of the learning system, we impose the following restrictions on the choice of the learning dynamics and the confidence levels.

**Assumption III.4:** There exists a stationary learner state $s^*$ such that $s^* = \mathcal{L}(s^*, \beta, w, v)$, for all $(\beta, w, v) \in \mathcal{I} \times \mathbb{R}^2$, and that from any initial state $s_0$, $\lim_{t \to \infty} s_t = s^*$, a.s.

**Assumption III.5:** The confidence dynamics $\beta_{t+1} = C(\beta_t)$ is chosen such that (i) $\sum_{t=0}^{\infty} \beta_t < \infty$ and (ii) $\lim_{t \to \infty} \log \beta_t = 0$, elementwise.

Assumption III.4 imposes that asymptotically, the learner settles down to some value which is no longer modified by additional data. It is natural to expect that in such a state, the learner has acquired perfect knowledge of the underlying transition kernel, and the ambiguity sets $A_{\beta_t}(s^*, w) \in \mathcal{S}$ have all converged to a singleton. However, this is not necessarily the case. For instance, the trivial case where $\mathcal{S} = \{s^*\}$ and $A_{\beta}(s, w) = \Delta_d$, $\forall(s, w) \in \mathcal{S} \times \mathcal{W}$ satisfies Assumption III.4, but under these conditions, no learning occurs, and in fact, a robust MPC scheme is recovered. In Section V-D, we will pose an additional assumption on the learning system, which excludes this case, but allows us to show consistency of the learning controller.

Assumption III.5 states that the probability of obtaining an ambiguity set that contains the true conditional distribution [expressed by (8)] increases sufficiently fast [Condition (i)]. This assumption will be of crucial importance in showing stability (see Section V-C). In addition, it places a lower bound on the convergence rate of the confidence levels [condition (ii)], which is crucial in establishing asymptotic consistency of the scheme (see Section V-D), since it will allow convergence of the ambiguity sets, as we discuss in Remark III.12. To fix ideas, we keep the following example in mind as a suitable choice for the confidence dynamics throughout this article.

**Example III.6 (Confidence dynamics):** A suitable family of sequences for the confidence levels satisfying Assumption III.5 (assuming $n_\beta > 1$ for simplicity) is obtained as

$$\beta_t = b(1 + t)^{-q}, \quad t \in \mathbb{N}$$

with parameters $0 < b \leq 1, q > 1$. Indeed, as $q > 1$, this sequence is summable, and furthermore, $\lim_{t \to \infty} -q \log(b(1 + t)) = 0$. Using (9), a straightforward calculation reveals that $\beta_t$ can be updated recursively $\beta_{t+1} = b^{-1/q}(2 + t) = b^{-1/q} + \beta_t^{1/q}$, and therefore, $\beta_{t+1} = b\beta_t \left( b^{1/q} + \beta_t^{1/q} \right)^{-q} =: C(\beta_t)$ Thus, it additionally satisfies the requirements of Assumption III.1.

The learner state $s_t$ will in most practical cases be composed of a data-driven estimator for the transition kernel and some parameter that is calibrated using the size of the ambiguity set, based on statistical information. In particular, we will focus on divergence-based ambiguity sets, defined as a ball (defined in some divergence) around an empirical estimate of the distribution. For the current setting concerning finitely supported distributions, two notable examples of such divergences are the total variation (TV) metric \cite{11, 48, 49} and the Kullback–Leibler (KL) divergence \cite{7}.

In the following section, we show that these divergences can be used to design a learning system satisfying our assumptions, and illustrate that from these cases, several other divergence-based ambiguity sets can be constructed.

---

**B. Divergence-Based Ambiguity Sets**

Our goal is to obtain for each mode $w$ of the Markov chain, a data-driven subset of the probability simplex, containing the $w$th row of the transition kernel $P$ with high probability. Given a sequence $\hat{w}_{[t, \ell]} \in \mathcal{W}^\ell$ of $\ell \in \mathbb{N}$ samples drawn from the Markov chain, $\ell$ individual datasets $\hat{W}_{t, \ell} := \{ \hat{w}_{k+1} \in \mathcal{W} \mid k \in [n_{\ell}], \hat{w}_{k} \mid w \in \mathcal{W}$, can be obtained by partitioning the set of observed transitions by the mode they originated in. As such, each $\hat{W}_{t, \ell}$ contains $\ell$ independent identically distributed (i.i.d.) draws from the distribution $P_t$. Ambiguity sets can now be constructed for each individual row $w$, using concentration inequalities based on the data in $\hat{W}_{t, \ell}$.

With this set-up, we now consider the following instance of a learning system.

**Definition III.7 (Empirical learner):** Let the learner state be composed as $s_t = (\text{vec} P_t, \gamma_t) \in \mathcal{S} = \Delta_d \times [0, 1]^d$, where $P_t$ denotes the empirical transition probability matrix at time $t$, i.e.,

$$P_{t,ij} = \begin{cases} \frac{1}{t} \sum_{w \in \hat{W}_{t,1}} 1_{w=j} & \text{if } t > 0 \\ \frac{1}{d} & \text{otherwise} \end{cases}$$

and $\gamma_t = (\frac{1}{t+1})_{i \in \mathcal{W}}$ is a vector containing the inverse of the mode-specific sample sizes.

For this instance of a learning system, we can now easily derive an explicit characterization of $L$.

**Example III.8 (Dynamics of the empirical learner):** The learner state $s_t$ is composed of $s_t = (\text{vec} P_t, \gamma_t)$. For the update of the empirical distribution $\hat{P}_t$, note that if $w_t \neq i$, then trivially, $\hat{P}_{t+1,i} = \hat{P}_{t,i}$. Otherwise, we may use the following well-known construction. Let $e_w \in \mathbb{R}^d$ denote the $w$th standard basis vector in $\mathbb{R}^d$, then

$$\hat{P}_{t+1,i} = \begin{cases} \frac{1}{t+1} \sum_{w \in \hat{W}_{t+1,i}} e_w & \text{if } w_t = i \\ \frac{1}{t+1} \left( \sum_{w \in \hat{W}_{t,i}} e_w + e_{w_{t+1,i}} \right) & \text{if } w_t = i \end{cases}$$

Since for all $w \in \mathcal{W}$, we define

$$L_{1,\ell}(\hat{P}_t, \gamma_t, w_t) := \begin{cases} (1 - \gamma_t,i) \hat{P}_{t,i} + \gamma_t,i e_{w_{t+1}} & \text{if } w_t = i \\ \hat{P}_{t,i} \end{cases}$$

Similarly, if $w_t = i$, $w_{t+1} \neq i$; otherwise, it follows from the definition of $\gamma_t,i$ that $\gamma_{t+1,i} = \frac{\gamma_t,i}{1 + \gamma_t,i}$, resulting in

$$L_{2,\ell}(\gamma_t, w_{t+1}) := \begin{cases} \frac{\gamma_t,i}{1 + \gamma_t,i} & \text{if } w_{t+1} = i \\ \frac{\gamma_t,i}{1 + \gamma_t,i} & \text{otherwise} \end{cases}$$

Concatenating (10) and (11), we obtain the Markovian update required by Assumption III.1

$$L(s_t, w_t, w_{t+1}) = \left( L_{1,\ell}(\hat{P}_t, \gamma_t, w_t) \right)_{i \in \mathcal{W}} \left( L_{2,\ell}(\gamma_t, w_{t+1}) \right)_{i \in \mathcal{W}}$$

Furthermore, this system satisfies Assumption III.4. Indeed, given ergodicity of the Markov chain (Assumption II.1), the

---

2For $n_\beta > 1$, the same construction can be repeated elementwise.

3The inversion results in simpler updates and renders $S$ robustly positive invariant, i.e., $s_t \in S \Rightarrow s_{t+1} \in S$. 

---
TABLE I

| Divergence | $\mathcal{D}(\hat{p}, p)$ | radius $r(m, \beta)$ | Conic representation |
|------------|--------------------------|----------------------|----------------------|
| TV         | $|p - \hat{p}|_1$         | $2\sqrt{rTV(m, \beta)}$ | Linear               |
| KL (KL)    | $\mathcal{D}_{KL}(\hat{p}, p)$ | $rKL(m, \beta)$ | Exponential           |
| JS (JS)    | $\frac{1}{2} \left( \mathcal{D}_{KL}(\hat{p}, \frac{\hat{p} + p}{2}) + \mathcal{D}_{KL}(\hat{p}, \frac{\hat{p} + p}{2}) \right)$ | $\frac{1}{2} rKL(m, \beta)$ | Exponential           |
| (Squared) Hemming (H) | $\sum_{i \in W} \left( \sqrt{p_i} - \sqrt{\hat{p}_i} \right)^2$ | $rKL(m, \beta)$ | Quadratic             |
| Wasserstein* (W) | $\min_{i,j \in W} \left\{ \sum_{s \in W} \Pi_{ij} K_{ij} | \Pi d = p, \Pi^+ d = \hat{p} \right\}$ | $\max_{i,j \in W} K_{ij} \sqrt{rTV(m, \beta)}$ | Linear               |

*Assume $W$ is a metric space. $K \in \mathbb{R}^{d \times d}$ is a symmetric distance kernel with $K_{ij} = \text{dist}(i, j), \forall i, j \in W$.  

Borel–Cantelli lemma [46, Th. 4.3] in conjunction with [50, Lemma 6] guarantees that with probability 1, there exists a finite time $T$, such that for all $t > T$ and for all $i \in W$, it holds that $t_i \geq ct$, where $c > 0$ is a constant depending on specific parameters of the Markov chain, and $t_i$ denotes the number of visits to mode $i$. That is, all modes are visited infinitely often, and as a result, both $\lim_{t \to \infty} \gamma_t = 0$ and $\lim_{t \to \infty} \tilde{P}_t = \hat{p}$, which are indeed fixed points of (10) and (11).

We can now associate with the newly defined empirical learner the following wide class of ambiguity sets, which take the form of a ball around the empirical estimate in some given statistical divergence.

Definition IV.3 (Divergence-based ambiguity set): Consider the empirical learner with state $s_t = (\text{vec} \tilde{P}_t, \gamma_t)$. We say that an ambiguity set $A_{\beta_t}(s_t, w)$ is a divergence-based ambiguity set if it can be expressed in the following form

$$A_{\beta_t}(s_t, w) := \{ p \in \Delta_\Delta | \mathcal{D}(\tilde{P}_t, p) \leq r(\gamma_{t,w}^{-1} - 1, \beta_t) \}, \forall w \in W,$$

where $\mathcal{D} : \Delta_\Delta \times \Delta_\Delta \to \mathbb{R}_+$ is some statistical divergence and $r : \mathbb{R}_+ \times [0, 1] \to \mathbb{R}_+$ is a given function that returns a radius, given a sample size and a confidence level.

Statistically meaningful values for the radius $r$ under different choices of divergences can be obtained using the following standard results.

Proposition III.10 (Concentration inequalities): Let $p \in \Delta_\Delta$ denote a distribution on the probability simplex and $\beta_t \overset{\text{m.i.d.}}{\sim} \frac{1}{m} \sum_{i=0}^{m-1} (1, w_{i+1})_1$, the empirical distribution based on $m$ i.i.d. draws $w_t \sim p$. Then, $\mathbb{P}\left[ \left( \frac{d}{2} \| p - \hat{p} \|_1 \right)^2 > r^{tv}(m, \beta) \right] \leq \beta$, with

$$r^{tv}(m, \beta) = \frac{d \log 2 - \log \beta}{2m}. \quad (12)$$

Similarly, it holds that $\mathbb{P}\left[ \mathcal{D}_{KL}(\hat{p}, p) > r^{KL}(m, \beta) \right] \leq \beta$, with

$$r^{KL}(m, \beta) = \frac{d \log m - \log \beta}{m}. \quad (13)$$

where $\mathcal{D}_{KL}(p, q) := \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i}$ denotes the KL divergence from $q$ to $p$.

The bound on the TV distance (12) is known as the Bretagnolle–Huber–Carol inequality [51, Th. A.6.6].

Remark III.11: Expression (13) for the KL radius is a well-known result from the field of information theory, obtained through the so-called method-of-types [52, 53]. A slight improvement can be obtained by replacing $d \log m$ by $\log \frac{(m+\epsilon-1)}{\epsilon-1}$.

Moreover, in [54], an even sharper result for (13) is derived. In fact, this improved concentration bound in the KL divergence was used in the same work to improve upon the TV concentration bound (12) for $\frac{m}{d} \ll 1$, using Pinsker’s inequality [55], which relates the TV distance between distributions $p, q \in \Delta_\Delta$ to the KL divergence as $\| p - q \|_1^2 \leq 2D_{KL}(p, q)$. Of course, these improved bounds can be readily used in practice to replace those in Theorem III.10. However, for the theoretical discussion, these modifications are inconsequential. For this reason, we opt to develop the ideas for the simpler, more commonly used forms.

Besides Pinsker’s inequality, there exist several other inequalities relating different statistical divergences (see for instance [56] for a comprehensive overview). Based on these relations, one can derive from Theorem III.10 several divergence-based ambiguity sets defined through other statistical divergences. For instance, since the squared Hellinger divergence is upper bounded by the KL divergence, (13) can be used as a radius for Hellinger divergence-based ambiguity sets. A summary of the resulting radii is provided in Table I. The rightmost column in this table refers to the conic representation of the induced ambiguity sets [cf., (23)], which determines the complexity of the resulting optimal control problems (see [74, Appendix A] for more details). Other works that have used these divergences (which belong to the class of $\phi$-divergences) for DR optimization are [57], [58], and [59]. In these works, however, the radii are either selected as a tuning parameter or calibrated using asymptotic arguments, leading to approximate ambiguity sets, which satisfy the coverage condition (8) only as the sample size tends to infinity. By contrast, the radii given in Table I are valid for any sample size.

We conclude the section by proposing a useful extension of the learner state in the case of divergence-based ambiguity sets.

Remark III.12 (Radius as part of the learner state): For divergence-based ambiguity sets, it is often convenient to augment the learner state $s_t = (\text{vec} \tilde{P}_t, \gamma_t)$ with the computed radii $r_{t,i} := r(\gamma_{t-1}^{-1} - 1, \beta_t)_{i \in W}$, for which the recursive update is obtained simply by composition of $r$ with the previously designed $\mathcal{L}$ and $C$. It can be easily verified using Theorem III.10 that this quantity also converges to the fixed point $\lim_{t \to \infty} r_{t,i} = 0 \forall i$. Indeed, Recall from Example III.8 that $\gamma_{t-1}^{-1} \sim t$ as $t \to \infty$. Using a radius function $r$ based on either (12) or (13), we obtain

$$\lim_{t \to \infty} r_{t,i} = \lim_{t \to \infty} r(t, \beta_t) = \lim_{t \to \infty} \frac{-\log \beta_t}{t} = 0$$

where the last equality follows from Assumption III.5.

IV. LEARNING MODEL PREDICTIVE CONTROL

Given a learning system satisfying Assumption III.1, we define the augmented state $y_t = (x_t, s_t, \beta_t) \in \mathcal{Y} := \mathbb{R}^{n_x} \times \mathcal{S} \times \mathbb{R}_+$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
\( T \), which evolves over time according to the dynamics

\[
y_{t+1} = f(y_t, w_t, u_t, w_{t+1}) := \left[ f(x_t, u_t, w_{t+1}) \right] C(\beta)
\]

with \( w_{t+1} \sim P_{w_{t+1}} \), for \( t \in \mathbb{N} \). Furthermore, it will be convenient to define the process \( z_t = (y_t, w_t) \in \mathcal{Z} := \mathcal{Y} \times \mathcal{W} \). Consequently, the objective is now to obtain a feedback law \( \kappa : \mathcal{Z} \to \mathbb{R}^{n_x} \). To this end, we will formulate a DR counterpart to the stochastic OCP (7), in which the expectation operator in the cost and the conditional probabilities in the constraint will be replaced by operators that account for ambiguity in the involved distributions.

### A. Ambiguity and Risk

In order to reformulate the cost function (6), we first introduce an ambiguous conditional expectation operator, leading to a formulation akin to the Markovian risk measures utilized in [32] and [60]. Consider a function \( \xi : \mathcal{Z} \times \mathcal{W} \to \mathbb{R} \), defining a stochastic process \( (\xi_t)_{t \in \mathbb{N}} = (\xi(z_t, w_{t+1}))_{t \in \mathbb{N}} \) on \( (\Omega, F, \mathbb{P}) \), and suppose that the augmented state \( z_t = z = (x, s, \beta, w) \) is given. Let \( \beta \in [0, 1] \) denote an arbitrary component of \( \beta \). The ambiguous conditional expectation of \( \xi(z, v) \), given \( z \), is then

\[
\tilde{\mathbb{E}}_{\beta, w}^{\beta}[\xi(z, v)] := \max_{p \in A_{\beta}(s, w)} \mathbb{E}_p[\xi(z, v) | z]
\]

Trivially, it holds that if the \( w \)th row of the transition matrix lies in the corresponding ambiguity set, i.e., \( P_w \in A_{\beta}(s, w) \), then

\[
\tilde{\mathbb{E}}_{\beta, w}^{\beta}[\xi(z, v)] \geq \mathbb{E}_{P_w}[\xi(z, v) | z] = \sum_{v \in \mathcal{W}} P_w \xi(z, v).
\]

Note that the function \( \tilde{\mathbb{E}}_{\beta, w}^{\beta} \) defines a coherent risk measure [34, Sec. 6.3]. We say that \( \tilde{\mathbb{E}}_{\beta, w}^{\beta} \) is the risk measure induced by the ambiguity set \( A_{\beta}(s, w) \).

A similar construction can be carried out for the chance constraints (5). We robustify the average value-at-risk with respect to the reference distribution, defining

\[
\tilde{\mathbb{V}}_{\beta, w}^{\beta, \alpha}[\xi(z, v)] := \max_{p \in A_{\beta}(s, w)} \mathbb{V}^{\alpha}_{\mathbb{R}^q}[\xi(z, v) | z] \leq 0.
\]

The function \( \tilde{\mathbb{V}}_{\beta, w}^{\beta, \alpha} \) in turn defines a coherent risk measure. Note that we have replaced the \( \mathbb{V}^{\alpha} \mathbb{R}^q \) parameter \( \alpha \) by \( \beta \). The reason for this is that the ambiguity set contains the true distribution only with high probability. Considering this fact, it is natural to expect that \( \alpha \) needs to be tightened to some extent in order to ensure that the original chance constraint remains satisfied. We make this precise in the following result.

**Proposition IV.1:** Let \( \beta, \alpha \in [0, 1] \) be given values with \( \beta < \alpha \). Consider the random variable \( s : \Omega \to \mathcal{S} \), denoting an (a priori unknown) learner state satisfying Assumption III.1, i.e., \( \mathbb{P}[P_w \in A_{\beta}(s, w)] \geq 1 - \beta \). If the parameter \( \alpha \) is chosen to satisfy \( 0 \leq \alpha \leq \frac{\alpha - \beta}{1 - \beta} \leq 1 \), then, for an arbitrary function \( g : \mathcal{Z} \times \mathcal{W} \to \mathbb{R} \), the following implication holds:

\[
\tilde{\mathbb{V}}_{\beta, w}^{\beta, \alpha}[g(z, v)] \leq 0, \text{ a.s. } \Rightarrow \mathbb{P}[g(z, v) \leq 0 | x, w] \geq 1 - \alpha.
\]

**Proof:** If \( \tilde{\mathbb{V}}_{\beta, w}^{\beta, \alpha}[g(z, v)] \leq 0 \), a.s., then (4) and (17) imply that

\[
\mathbb{P}[g(z, v) \leq 0 | x, w, P_w \in A_{\beta}(s, w)] \geq 1 - \alpha, \text{ a.s.}
\]

Therefore,

\[
\mathbb{P}[g(z, v) \leq 0 | x, w] \geq \mathbb{P}[g(z, v) \leq 0 | x, w, P_w \in A_{\beta}(s, w)] \mathbb{P}[P_w \in A_{\beta}(s, w)] \geq (1 - \alpha)(1 - \beta).
\]

Requiring that \((1 - \alpha)(1 - \beta) \geq (1 - \alpha) \) then immediately yields the sought condition.

Note that the implication (18) in Proposition IV.1 provides an a priori guarantee, since the learner state is considered to be random. In other words, the statement is made before the data are revealed. Indeed, for a given learner state \( s \) and mode \( w \), the ambiguity set \( A_{\beta}(s, w) \) is fixed, and therefore, the outcome of the event \( E = \{ P_w \in A_{\beta}(s, w) \} \) is determined. Whether (18), then, holds for these fixed values, depends on the outcome of \( E \). This is naturally reflected through the abovementioned condition on \( \alpha \), which implies that \( \alpha \leq \beta \), and thus tightens the chance constraints that are imposed conditioned on a fixed \( s \). Hence, we account for the possibility that the ambiguity set may not include the conditional distribution for this particular \( s \). This tightening can be mitigated by decreasing \( \beta \), at the cost of a larger ambiguity set. A more detailed study of this tradeoff is left for future work.

### B. Distributionally Robust Model Predictive Control

We are now ready to describe the DR counterpart to OCP (7). Consider a given augmented state \( z = (x, s, \beta, w) \in \mathcal{Z} \). Hereafter, we will assume that \( \beta = (\beta, \beta) \), where component \( \beta \) is related to the cost function and \( \beta \) is reserved for the constraints.

We use (17) to define the DR set of feasible inputs \( \hat{U}(z) \) in correspondence to (5), as

\[
\hat{U}(z) := \{ u \in U : \tilde{\mathbb{E}}_{\beta, w}^{\beta}[g(x, u, w, v)] \leq 0 \}.
\]

**Remark IV.3:** The parameter \( \alpha \) remains to be chosen in relation to the confidence levels \( \beta \) and the original violation rates \( \alpha \). In light of Proposition IV.1, \( \alpha = \frac{\alpha - \beta}{1 - \beta} \) yields the least conservative choice. This choice is valid as long as it is ensured that \( \beta < \alpha \).

Using (15), we express the DR cost of a policy \( \pi = (\pi_k)_{k=0}^{N-1} \)

\[
\hat{V}_N(z) := \ell(x_0, u_0, w_0) + \rho^{\alpha_0}_{s_0, w_0} \left[ \ell(x_1, u_1, w_1) + \rho^{\alpha_1}_{s_1, w_1} \left[ \cdots + \rho^{\alpha_{N-2}}_{s_{N-2}, w_{N-2}} \left[ \ell(x_{N-1}, u_{N-1}, w_{N-1}) + \rho^{\alpha_{N-1}}_{s_{N-1}, w_{N-1}} [\hat{V}_I(x_N, s_N, \beta_N, w_N)] \right] \right] \right] \quad (20)
\]

where \( z_0 = z, z_{k+1} = \hat{f}(z_k, u_k, w_{k+1}), \) and \( u_k = \pi_k(z_k) \), for all \( k \in [0, N-1] \). In Section V, conditions on the terminal cost \( \hat{V}_I : \mathcal{Z} \to \mathbb{R}_+ : (x, s, \beta, w) \mapsto \hat{V}_I(x, w) + \delta(z, s, \beta, w) \) and its domain are provided in order to guarantee recursive feasibility and stability of the MPC scheme defined by the following OCP.

**Definition IV.3 (DR-OCP):** Given an augmented state \( z \in \mathcal{Z} \), the optimal cost of the distributionally robust optimal control problem (DR-OCP) is

\[
\hat{V}_N(z) = \min_{\pi} \hat{V}_N^\pi(z) \quad (21a)
\]

subject to

\[
(x_0, s_0, \beta_0, w_0) = z, \quad \pi = (\pi_k)_{k=0}^{N-1} \quad (21b)
\]
\[ z_{k+1} = (f(z_k, \pi_k(z_k), w_{k+1}), w_{k+1}) \quad (21c) \]

\[ \pi_k(z_k) \in \hat{U}(z_k) \quad \forall u \in [0, k] \in W^k \quad (21d) \]

for all \( k \in \mathbb{N}_{[0, N-1]} \). We denote by \( \hat{\Pi}_N(z) \) the corresponding set of minimizers.

**Remark IV.4:** Note that the definition of \( \hat{V}_t \) implicitly imposes the terminal constraint \( z_N \in X_t \), a.s.

We now define the learning MPC law analogously to the stochastic case as

\[ \hat{\pi}_N(z) = \hat{\pi}_N^*(z) \quad (22) \]

where \( (\hat{\pi}_N^*(z))_{k=1}^{N-1} \in \hat{\Pi}_N(z) \). At every time \( t \), the learning MPC scheme, thus, consists of repeatedly

1. solving (21) to obtain a control action \( u_t = \hat{\pi}_N(z_t) \) and applying it to the system (1);
2. observing the outcome of \( w_{t+1} \in W \) and the corresponding next state \( z_{t+1} = f(x_t, u_t, w_{t+1}) \);
3. updating the learner state \( s_{t+1} = \hat{L}(s_t, w_{t+1}) \) and the confidence levels \( \beta_{t+1} = C(\beta_t) \), gradually decreasing the size of the ambiguity sets.

Note that in its general form, (21) is a nonsmooth, infinite-dimensional optimization problem. However, provided that the involved risk measures are *conic risk measures* (as defined by Definition IV.5), problem (21) can be reformulated as a finite-dimensional, smooth nonlinear program.

**Definition IV.5 (Conic risk measure [44]):** We say that an ambiguity set \( A \subseteq \Delta_d \) is conic representable if it can be written in the form

\[ A = \{ p \in \Delta_d \mid \exists \nu : E p + F \nu \leq \kappa b \} \quad (23) \]

with matrices \( E \) and \( F \), vector \( b \) of suitable dimensions, and a proper cone \( \kappa \). The coherent risk measure induced by a conic representable ambiguity set is called a conic risk measure.

Since ambiguity sets inducing coherent risk measures are convex by construction, many classes of ambiguity sets can be represented using conic inequalities. For completeness, we state the conic representations for the ambiguity sets summarized in Table I, as well as the reformulation of (21) in [74, Appendix A and B], respectively.

**V. THEORETICAL ANALYSIS**

**A. Dynamic Programming**

To facilitate the theoretical analysis of the proposed MPC scheme, we follow an approach similar to [32] and represent (21) as a dynamic programming recursion. We define the Bellman operator \( T \) as \( T(\hat{V})(z) := \min_{s(x, u, v) \in [0, k]} f(x, u, v) + \rho^\beta \rho_{w,v}[\hat{V}(\hat{z}, u, v), v] \), where \( \rho^\beta \rho_{w,v}[\hat{V}(\hat{z}, u, v), v] \) is the corresponding set of minimizers. The optimal cost \( \hat{V}_N(z) \) is obtained through the iteration

\[ \hat{V}_k = T \hat{V}_{k-1}, \quad \hat{V}_0 = \hat{V}_r, \quad k \in \mathbb{N}_{[1, N]} \]

Similarly, \( \hat{Z}_k := \text{dom} \hat{V}_k \) is given recursively by

\[ \hat{Z}_k = \{ z \mid \exists u \in \hat{U}(z) : (\hat{f}(z, u, v), v) \in \hat{Z}_{k-1} \quad \forall v \in W \} \]

Now consider the stochastic closed-loop system

\[ y_{t+1} = \hat{f}_N(z_t, w_{t+1}) := \hat{f}(z_t, \hat{\pi}_N(z_t), w_{t+1}) \quad (25) \]

where \( \hat{\pi}_N(z_t) \in S(\hat{V}_{N-1}(z_t)) \) is an optimal control law obtained by solving the DR-OCP of horizon \( N \) in receding horizon.

**B. Constraint Satisfaction and Recursive Feasibility**

In order to show the existence of \( \hat{\pi}_N \in S(\hat{V}_{N-1}) \) at every time step, Proposition V.4 will require that \( \hat{X}_t \) is a RCI set. We define robust control invariance for the augmented control system under consideration as follows.

**Definition V.1 (Robust control invariance):** A set \( R \subseteq Z \) is an RCI set for the system (14) if for all \( z \in R \), \( \exists u \in \hat{U}(z) \) such that \( (f(z, u, v), v) \in R, \forall v \in W \). Similarly, \( R \) is a robust positive invariant (RPI) set for the closed-loop system (25) if for all \( z \in R \), \( (\hat{f}_N(z, v), v) \in R, \forall v \in W \).

Since \( \hat{U} \) consists of conditional risk constraints, our definition of robust invariance provides a DR counterpart to the notion of *stochastic* robust invariance in [61]. This notion is less conservative than the following, more classical notation of robust invariance.

**Definition V.2 (Classical robust control invariance):** A set \( R_x \subseteq \mathbb{R}^n \times \mathcal{W} \) is an RCI set for system (1) in the classical sense if for all \( (x, w) \in R_x \)

\[ \exists u : g(x, u, w, v) \leq 0, (f(x, u, v), v) \in R_x(v), \forall v \in W. \]

In fact, for any set \( R_x \) as in Definition V.2, the set \( R_x \times \mathcal{S} \times \mathcal{L} \) is covered by Definition V.1, as illustrated in Example V.3.

On the other hand, our notion of robust control invariance is stricter than that of *uniform control invariance* considered in [32], which only requires successor states to remain in the invariant set for modes \( v \) in the *cover* of the given mode \( w \), i.e., the set of modes \( v \) for which \( P_{w,v} > 0 \). This flexibility is not available in the current setting, as the transition kernel is assumed to be unknown, so the cover of a mode cannot be determined with certainty.

**Example V.3 (Classical robust invariant set):** Suppose that the terminal constraint set \( \hat{X}_N(w) := \{ x \mid (x, w) \in \hat{X}_N \} \). Then, if \( \hat{X}_t \) is chosen such that \( \hat{X}_t(w) := \{ y \mid (y, w) \in \hat{X}_t \} = \hat{X}_t(w) \times \mathcal{S} \times \mathcal{L} \), \( \hat{X}_t(w) \) is an RCI set for the augmented system (14) according to Definition V.1. Indeed, since \( \hat{V}_N \circ P_{w,v}[g(x, u, w, v) \leq \max_{x,v} g(x, u, w, v) \forall \alpha \in [0, 1] \) and \( p \in \Delta_d \), (26) implies that for all \( z \in \hat{X}_N \), there exists \( u \in \hat{U}(z) \), such that \( f(z, u, v) \in \hat{X}_N(v) \).

**Proposition V.4 (Recursive feasibility):** If \( \hat{X}_r \) is an RCI set for (14), then (21) is recursively feasible. That is, feasibility of DR-OCP (21) for some \( z \in \mathcal{Z} \), implies feasibility for \( z^+ = \hat{f}(x, u, v, v), \forall v \in \mathcal{W}, N \in \mathbb{N}_{>0} \).

**Proof:** The proof follows from a straightforward inductive argument on the prediction horizon \( N \). We first show that if \( \hat{X}_r \) is RCI, then so is \( \hat{Z}_r \). This is done by induction on the horizon \( N \) of the OCP.

**Base case (\( N = 0 \)):** Trivial, since \( \hat{Z}_0 = \hat{X}_r \).

**Induction step (\( N \Rightarrow N + 1 \)):** Suppose that for some \( N \in \mathbb{N} \), \( \hat{Z}_N \) is an RCI set for (14). Then, by definition of \( \hat{Z}_{N+1} \), there exists for each \( z \in \hat{Z}_{N+1} \), a nonempty set \( \hat{U}_N(z) \subseteq \hat{U}(z) \) such that for every \( u \in \hat{U}_N(z) \), and \( v \in \mathcal{W} \), it holds that \( z^+ \in \hat{Z}_N \), where \( z^+ = \hat{f}(z, u, v) \). Furthermore, the induction hypothesis \( \hat{Z}_N \) is an RCI set for the system (25) implies that there also exists a \( u^+ \in \hat{U}(z^+) \) such
that \( \hat{f}(z^+, u^+, v^+) \in \mathcal{Z}_N(v^+) \), \( \forall v^+ \in W \). Therefore, \( z^+ \) satisfies the conditions defining \( \hat{Z}_{N+1} \). In other words, \( \hat{Z}_{N+1} \) is RCI.

The claim follows from the fact that for any \( N > 0 \) and \( z \in \hat{Z}_N \), \( u = \hat{r}_N(z) \in S(\hat{V}_{N-1}(z)) \subseteq \hat{U}_{N-1}(z) \), as any other choice of \( u \) would yield infinite cost in the definition of the Bellman operator.

**Corollary V.5 (Chance constraint satisfaction):** If the conditions for Proposition V.4 hold, then by Proposition IV.1, the stochastic process \( (z_t)_{t \in \mathbb{N}} = (x_t, s_t, \beta_t, w_t)_{t \in \mathbb{N}} \) satisfying dynamics (25) satisfies the nominal chance constraints

\[
\mathbb{P}[g(x_t, \hat{r}_N(z_t), w_{t+1}) > 0 \mid x_t, w_t] < \alpha
\]

a.s., for all \( t \in \mathbb{N} \).

We conclude this section by emphasizing that although the MPC scheme guarantees closed-loop constraint satisfaction, it does so while being less conservative than a fully robust approach, which is recovered by taking \( A_\beta(s, w) = \Delta_\beta \) for all \( (s, w, \beta) \in \mathcal{S} \times W \times [0, 1] \). It is apparent from (17) and (19) that for all other choices of the ambiguity set, the set of feasible control actions will be larger (in the sense of set inclusion).

**C. Stability**

In this section, we will provide sufficient conditions on the control setup under which the origin is MSS for (25), i.e., \( \lim_{t \to \infty} \mathbb{E}[\|x_t\|^2] = 0 \) for all \( x_0 \) in some specified compact set containing the origin.

Our main stability result, stated in Theorem V.7, hinges in large on the following lemma, which relates risk-square stability [32, Th. 6] of the origin for the autonomous system (25) (with respect to a statistically determined ambiguity set) to stability in the mean-square sense (with respect to the true distribution).

**Lemma V.6 (DR MSS condition):** Suppose that Assumption III.5 holds and that there exists a nonnegative, proper function \( V : \hat{Z} \to \mathbb{R}_+ \), such that

1. \( \text{dom } V \) is RPI for (25) and \( \text{dom } V(\cdot, s, \beta, w) \) is compact and contains the origin for all \( (s, \beta, w) : \text{dom } V(\cdot, s, \beta, w) \neq \emptyset \);
2. \( \rho_{s,w}^2[V(\hat{r}_N, z, v)] - V(z) \leq -\varepsilon \|x\|^2 \), for some \( c > 0 \), for all \( z \in \text{dom } V \);
3. \( V \) is uniformly bounded on its domain. Then, \( \lim_{t \to \infty} \mathbb{E}[\|x_t\|^2] = 0 \) for all \( x_0 \in \text{dom } V \), where \( (z_t)_{t \in \mathbb{N}} = (x_t, s_t, \beta_t, w_t)_{t \in \mathbb{N}} \) is the stochastic process governed by dynamics (25).

**Proof:** See Appendix B.

**Theorem V.7 (MPC stability):** Suppose that Assumptions II.3 and III.5 are satisfied and the following statements hold:

i) \( TV_t \leq \tilde{V}_t \);

ii) \( \varepsilon \|x\|^2 \leq e(x, u, w) \) for some \( c > 0 \), for all \( z = (x, s, \beta, w) \in \text{dom } \tilde{V} \) and all \( u \in \tilde{U}(z) \);

iii) \( \hat{V}_N \) is locally bounded on its domain.

Then, the origin is MSS for the MPC-controlled system (25), over all RPI sets \( \hat{Z} \subseteq \text{dom } \hat{V}_N \) such that for all \( (s, \beta, w) : (x, s, \beta, w) \in \hat{Z} \), the projection \( \{x \mid (x, s, \beta, w) \in \hat{Z} \} \) is compact and contains the origin.

**Proof:** The proof is along the lines of that of [32, Th. 6] and shows that \( \hat{V}_N \) satisfies the conditions of Lemma V.6. Details are given in Appendix B.

The results in this section indicate that after an appropriate choice of the learning system, the thusly defined risk measures can be used to design and MPC controller using existing techniques (e.g., those presented in [32]). Corresponding stability guarantees (assuming known transition probabilities) then translate directly into stability guarantees under an ambiguously estimated transition kernel.

**D. Out-of-Sample Bounds and Consistency**

We now turn our attention to analyzing the value function of the DR-OCP in relation to the nominal (stochastic) OCP. We will show that under quite general assumptions, the former provides an upper bound to the latter with high probability (Theorem V.9). Furthermore, under appropriate constraint qualifications, we will show that the optimal value of the DR-OCP converges to that of the nominal problem as the sample size increases, see Theorem V.13. In the particular case where the constraints do not depend on the distribution, we can relax the constraint qualification to obtain a similar result. We include this as a separate statement, as it permits a more direct and illustrative proof using dynamic programming.

Given an arbitrary state-mode pair \((x, w)\), initial value of the learning state \(s_0\), and confidence \(\beta_0\), the stochastic process defined by the optimal value of the DR-OCP (21), i.e., \( \hat{V}_N(t, x, w) := \hat{V}_N(x, s_t, \beta_t, w), t \in \mathbb{N} \), serves as a sequential approximation of the optimal value \(V_N(x, w)\) of the horizon-\(N\) nominal OCP (7). This section will establish sufficient conditions under which \(\hat{V}_N(t)\) bounds \(V_N\) from abovementioned, and for which it converges to \(V_N\) almost surely—a property which we refer to as asymptotic consistency. The former guarantee will provide a performance certificate in the sense that the true optimal cost (under full knowledge of the distribution) will be no worse than the cost predicted by solving its DR counterpart. Of course, this guarantee is also provided by a robust (minimax) scheme (obtained by taking \(A_\beta \equiv \Delta_\beta, \forall \beta\)). However, such an approach is nonadaptive and, therefore, lacks consistency. On the other hand, a sample-average approximation (in which the ambiguity set is replaced by a singleton containing only the empirical distribution) may under similar conditions be consistent, but it provides no safety guarantees/performance bounds.

In the following, we denote \(\hat{X}_t(w) = \{x \mid (x, w) \in \hat{X}_t\}\) and similarly \(\hat{X}_t(w) = \{y \mid (y, w) \in \hat{X}_t\}\). We will also pose the following assumptions in the remainder of the section.

**Assumption V.8:**

i) The risk levels \(\tilde{\alpha}_t\) are chosen according to the upper bound of Proposition IV.1, i.e., \(\tilde{\alpha}_t = \frac{\alpha - \beta_t}{1 - \beta_t}\) and \(\beta_t < \alpha \leq 1\).

ii) \(\tilde{X}_t\) is constructed in relation to the original problem such that for all \((w) \in W\), \(\hat{X}_t(w) = \hat{X}_t(w) \times \mathcal{S} \times \mathcal{T}\), and \(\hat{X}_t\) is RCI for system (1) in the sense of Definition V.2.

**Theorem V.9 (Performance guarantee):** Suppose that Assumption V.8 holds. Then, for any initial learner state \(s_0 = s \in \mathcal{S}\) and any initial confidence level \(\beta_0 = \beta \in \mathcal{T}\),

i) the value function of the DR-OCP of horizon \(N \geq 0\) asymptotically upper bounds the true value function. That is

\[
\mathbb{P}[\hat{V}_N(t)(x, w) \geq V_N(x, w) \forall (x, w) \in \text{dom } V_N] \geq 1 - \gamma(t)
\]

(27)

for all \(t \in \mathbb{N}\), with \(\gamma(t) = d \sum_{k=0}^{t-N} \|\beta_k\|_1\).
ii) If, furthermore, $A_\beta$ is selected such that
\[
\beta' \leq \beta \Rightarrow A_\beta(s, w) \subseteq A_{\beta'}(s, w) \quad \forall (s, w) \in S \times W
\]
then, (27) holds with $\gamma_N^{(t)} = d \sum_{k=1}^{N} \|\beta_k\|_\infty$.

**Proof:** See Appendix B.

Theorem V.9 guarantees that with high probability, the DR value function provides an upper bound for the value function under full knowledge of the distribution. The corresponding violation rate $\gamma(N)$ can be tuned using the user-specified confidence levels $\beta_t$.

Note that the violation rate $\gamma(N)$ increases with the prediction horizon. This is to be expected, since we essentially require the ambiguity set to contain the true switching distribution for all predicted time steps, which becomes increasingly difficult as the horizon length increases. However, due to the summability of the confidence levels $\beta_t$ (cf., Assumption III.5), the violation rate $\gamma(N)$ will converge to a finite value as $N \to \infty$. Similarly, as $t \to \infty$ for fixed $N$, $\gamma(N)$ converges to zero at a summable rate. We will use this fact in Corollary V.10 to obtain a stronger guarantee asymptotically.

Before stating the asymptotic extension of Theorem V.9, we briefly highlight the sharper bound for $\gamma(N)$ stated in Theorem V.9-(ii). This result requires that for a given learner state $s$, the size of the ambiguity set scales monotonically with the required confidence level. This is satisfied for the described divergence-based ambiguity sets in Table I. Indeed, the center of the divergence balls are given by the empirical distribution and, therefore, independent of the confidence level $\beta$. The radii, by proposition III.10, are monotone decreasing functions of $\beta$. Thus, the intersection of a collection of such ambiguity sets is equal to the ambiguity set with the largest value of $\beta$ (and thus, the smallest radius).

**Corollary V.10:** Under the same conditions as Theorem V.9, we have with probability one that
\[
\hat{V}_N^{(t)}(x, w) \geq V_N(x, w) \text{ for all sufficiently large } t
\]
for all $(x, w) \in \text{dom } V_N$.

**Proof:** For fixed $(x, w) \in \text{dom } V_N$, Theorem V.9 guarantees that $P[\hat{V}_N^{(t)}(x, w) < V_N(x, w)] \leq \gamma(N)$, where due to Assumption III.5, $\sum_{t=0}^{\infty} \gamma(N) = d \sum_{k=0}^{N} \sum_{t=0}^{\infty} \|\beta_{t+k}\|_\infty < \infty$. The claim then follows from the Borel-Cantelli Lemma [46, Th. 4.3].

Having established a performance bound on the true cost, we will now demonstrate the consistency of the method, starting with the special case where the constraints are independent of the learner state (Theorem V.12), before tackling the general case in Theorem V.13. To this end, we make the following assumption on the learner state and the corresponding ambiguity set.

**Assumption V.11 (Ambiguity decrease):** There exists a sequence $\{\delta_t\} \in \mathbb{R}$ with $\lim_{t \to \infty} \delta_t = 0$, such that
\[
\sup_{p, q \in A_{\delta_t}(x_t, w)} \|p - q\|_1 \leq \delta_t \quad \text{a.s.} \quad \forall (x_t, w) \in W \quad \forall t \in \mathbb{N}_{[1, n_0]}
\]

Assumption V.11 states that the ambiguity sets “shrink” to a singleton with probability one. Since the ambiguity is expected to decrease as more information is observed, this is a rather natural assumption, which is satisfied by most classes of ambiguity sets, such as the ones discussed in Section III (cf., Remark III.12).

**Theorem V.12 (Asymptotic consistency with hard constraints):** Suppose that all constraints are hard constraints, i.e., $\alpha = 0$, so that $U(z) = U(x, w)$ for all $z = (x, s, \beta, w)$. Then, for any state-mode pair $(x, w) \in \text{dom } V_N$, any initial learner state $s_0 = s \in S$ and any initial confidence level $\beta_0 = \beta \in \mathbb{I}$, the optimal cost of the DR-OCP of horizon $N \geq 0$ almost surely converges from above to the true optimal cost. That is, with probability one
\[
\lim_{t \to \infty} \hat{V}_N^{(t)}(x, w) = V_N(x, w)
\]
for all $(x, w) \in \text{dom } V_N$.

**Proof:** See Appendix B.

In the more general case, where aside from the cost, also the constraints are probabilistic and, therefore, dependent on the learner state, where some additional assumptions on the problem ingredients are required.

**Theorem V.13 (Asymptotic consistency under chance constraints):** Let $s^* \in S$ denote a stationary learner state (cf., Assumption III.4) and suppose that for a given state-mode pair $(x, w) \in \text{dom } V_N$, the following hold:

i) assumption V.8 holds, and $A_\beta(s, w)$ is closed and convex;

ii) the costs $\ell(\cdot, \cdot, w), V(\cdot, \cdot, w)$, constraints $g(\cdot, \cdot, w, v)$, and dynamics $f(\cdot, \cdot, w, v)$ are continuously differentiable;

iii) the ambiguity set $A_\beta(s, w)$ is conic representable with convex cone $\mathcal{K}$ and parameters $E_w(s, \beta), F_w(s, \beta)$, and $b_w(s, \beta)$ that depend smoothly on $s$ and $\beta$;

iv) Robinson’s constraint qualification [62, Def. 2.86] holds for (74, eq. (37)), for initial states $(x, w), s^* = s^*, \beta^* = 0$.

Then, $\lim_{t \to \infty} \hat{V}_N^{(t)}(x, w) = V_N(x, w), \text{ a.s.}$

**Proof:** The proof consists of showing that an equivalent reformulation of the problem, namely (74, eq. (37)), satisfies the conditions for [62, Prop. 4.4]. The claim then follows directly. We omit the details here in the interest of space, and refer to the technical report [74] for the full proof.

We conclude this section with a few brief remarks regarding the conditions of Theorem V.13. First, we note that using the learning system described in Section III (including the ambiguity radius as part of the learner state as suggested in Remark III.2), Condition (iii) is satisfied for all divergence-based ambiguity sets considered in Table I. Indeed, in the conic formulations provided in [74, Appendix A], we find that in all cases, the empirical distribution and the ambiguity radius enter linearly in the final conic form of the constraints. Second, we remark that Robinson’s constraints qualification [Condition (iv)] can be regarded as a generalization of the more well-known Mangasarian–Fromowitz constraint qualification [62, eq. 2.191] (see also [63, Prop. 3.3.8] or [64, 4.10]), which is in turn a generalization of the linear independence constraint qualification. It is a very common regularity assumption, ensuring several useful properties, such as boundedness of Lagrange multipliers. Of main importance for the purpose of proving consistency under probabilistic constraints, however, is that it provides metric regularity of the (now parametric) feasible set, which implies that the distance from the feasible set can upper bounded by a multiple of the constraint violation.
Fig. 1. Radii of the ambiguity sets versus sample size $t$. The shaded area delineate the empirical $\beta_t$ upper and lower quantiles of $D(\hat{p}_t, p)$ for different divergences $D$, computed over 200 Monte Carlo runs. The dashed lines represent the theoretical upper bounds given in Table I. $p_i = 1/d \forall i \in N_{d-1}$, $d = 5$ and $\beta_t = e^{-d(t+1)}/2$.

VI. ILLUSTRATIVE EXAMPLES

A. Ambiguity Sets

To illustrate the concentration inequalities provided in Theorem III.10 and Table I, we select a sequence of confidence levels $\beta_t = e^{-d(t+1)}$ satisfying summability (Assumption III.5), and we plot the radii corresponding to the considered divergences as a function of the sample size $t$ (see Fig. 1). For comparison, we recursively estimate compute the empirical estimate $\hat{p}_t$ of a fixed probability vector $p \in \Delta_d$ and plot the empirical upper and lower $\beta_t$-quantile of $D(\hat{p}_t, p)$ over 200 Monte Carlo runs. For the Wasserstein distance, a quadratic kernel $K_{uv} = (w-v)^2 \forall u, v \in N_{d-1}$ was used. For all divergences, the given bounds provide reasonable approximations; but in particular, we note that the TV bound is almost tight. Furthermore, it only requires linear constraints in its conic representation (23), making it an attractive choice in terms of both statistical and computational complexity.

B. Distributionally robust MPC

We consider a Markov jump linear system

$$x_{t+1} = A(w_{t+1})x_t + B(w_{t+1})u_t,$$

(31)

with $A(w) = \begin{bmatrix} 1 + \frac{w-1}{0.01} & 0.01 \\ 1 + 2.5\frac{w-1}{d} \end{bmatrix}$, $B(w) = I, w \in N_{d-1}$.

The state $x_t \in \mathbb{R}^2$ of this system, inspired by [66], models the deviation of temperatures from some nominal value of two adjacent servers in a data center. The actuators $u_t \in \mathbb{R}^2$ correspond to the amount of heating ($u_t > 0$) or cooling ($u_t < 0$) applied to the corresponding machines. The mode $w$ models the load on the servers. If $w = 1$, the system is idle and no heat is generated. If $w = d$, then the processors are fully occupied and a maximum amount of heat is added to the system. Note that the second server generates more heat under increasing loads.

The true-but-unknown transition probabilities are computed as $P_{i,j} = e^{-(j-i)/2}/\sum_{w=1}^{d} e^{-(w-i)/2} \forall i, j \in N_{d-1}$.

As in [66], we will use a mode-independent quadratic cost $\ell(x, u, w) = \|x\|^2 + 10\|u\|^2$.

We impose hard constraints $-1 \leq u \leq 1$ on the actuation and (nominally) impose robust chance constraints

$$AV_{\alpha}x_{t+1} \in \{H_t: x_{t+1} - h_t | x_t | x_t \leq \alpha \} \quad \text{with} \quad H_t = \begin{bmatrix} 1_n \\ 1_n \end{bmatrix}, h_t = \begin{bmatrix} 1_n \\ 0.5 \end{bmatrix}$$

for all $t \in N_{d-1}$, and $\alpha = 0.19$. Hence, in this example, we have $\phi_t(x, u, w, v) = H_t(A(v)x + B(v)u) - h_t$.

We compute stabilizing terminal ingredients offline using standard techniques from robust control. We compute a robust quadratic Lyapunov function $V_t(x) = x^TQ_t x$ along with a local linear control gain $K_t$ such that $V_t((A(w)+B(w))K)x \leq -\ell(x, Kx), \forall w \in W$ by solving a linear matrix inequality (LMI) as in [67]. The RCI terminal set $X_\ell$ is computed as the level set $X_\ell = \text{lev}_{\leq \ell} V_t$, where $\ell = \min \{h_t/\|Q^{-1/2}H_t\|_2\}$ is the largest value such that $\text{lev}_{\leq \ell} V_t$ lies inside the polyhedral set (in $R^{m_t}$) $H(A(w)+B(w))K \leq h \forall w \in W$.

For the DR controllers in the following, we choose confidence levels $\beta_t = (\beta_1, \beta_2)$ with $\beta_1 = \beta_2 = 0.19 \epsilon^{-2} < \alpha$ for the cost and the constraints, respectively, ensuring that Assumption III.5 is satisfied. For simplicity, we use identical confidence levels $\beta_t$ for all the constraints.

We compare the proposed DR-MPC controller with 1) the (nominal) stochastic MPC controller [see (7)], which we call omniscient as it has access to the true transition matrix $P$, 2) the robust MPC controller, obtained by solving (21), taking the ambiguity set $A_{\beta_2}(s, w) = A_{\beta_2}(s, w) = \Delta_d$ to be the entire probability simplex, regardless of the mode or learner state. Both the LMIs involved in the offline computation of the terminal ingredients and the online risk-averse optimal control problem (21) are solved using MOSEK [68] through the CVXPY [69] interface.

We fix the number of modes to $d = 3$, and take $N = 5$. All computations were performed on an Intel Core i7-7700 K CPU at 4.20 GHz.

1) Timings: To obtain an indication of the comparative computational burden of the different divergence-based ambiguity sets under consideration, we solve the described DR-OCP using the considered divergences 10 times each, for random initial states. Table II reports the average and maximum observed solver time. As expected, the TV and Wasserstein divergence result require the least amount of time, as they introduce only linear constraints. The Hellinger divergence, which introduces second-order cone constraints, results in slightly longer run times. The KL and JS divergence both introduce exponential cone constraints, resulting in the most computationally demanding OCPs.

2) Closed-Loop Simulation: Motivated by previous experiments, we next select the TV ambiguity set, and perform a more extensive closed-loop simulation. Fixing the initial state at $x = [0.5 \ 0.5]^T$, we perform 50 Monte Carlo simulations of the described MPC problems for 30 steps. As the simulation
time is rather short, we initialize the DR controller with 10 and 100 offline observations of the Markov chain to obtain more interesting comparisons. Hence, the simulation in the following essentially compares the controller responses after a sudden disturbance after 10 and 100 time steps. All considered controllers are recursively feasible and mean-square stabilizing by construction. By the nature of the problem set-up, the optimal behavior is to just barely stabilize the system with minimal control effort. However, the larger the uncertainty on the state evolution, the more the controller is forced to drive the states further away from the constraint boundary, leading to larger control actions and, consequently, larger costs.

This behavior can be observed in Figs. 2 and 3. Fig. 2 shows the controls and states over time and Fig. 3 presents the distribution of the closed-loop costs (sum of the stage costs over the simulation time). In the first time step, the robust controller takes the largest step, driving the state the furthest from the constraint boundary. As illustrated in Fig. 2 (right), this is particularly pronounced for the second component of the state vector, as it is more sensitive to the mode [cf., (31)]. The omniscient stochastic MPC, by contrast, has perfect knowledge of the transition probabilities, and by consequence is able to more slowly drive the state to the origin, reducing the control effort considerably. The DR controller naturally “interpolates” between these behaviors. Initially, it performs only marginally better than the robust controller (due to the very limited number of online learning steps). As it gets access to increasing sample sizes, however, it gradually approximates the behavior of the omniscient controller, while guaranteeing satisfaction of the constraints throughout.

3) Asymptotic consistency: To illustrate the consistency results from Section V-D, we fix the initial state-mode pair \( x_0 = [0.25 \ 0.25]^T, w_0 = 1 \) and recompute the solution to problem (21) to obtain \( \tilde{V}^{(t)} := \tilde{V}^{(t)}_N (x_0, w_0) \) for increasing sample sizes \( t \).

For comparison, we compute 1) the true value \( V^* := V_N (x_0, w_0) \) by solving the stochastic MPC problem (7), using the true transition probabilities, and 2) the robust value function \( V^r \), obtained by solving (21), taking the ambiguity set \( A_{\beta}(s, w) = \Delta_{\beta} \) to be the entire probability simplex, regardless of the mode or learner state.

Fig. 4 shows the relative difference between the DR value \( \tilde{V}^{(t)} \) and the true value \( V^* \) for the different statistical divergences. At very low sample sizes, the DR controllers achieve the same cost as the robust controller. However, as more data are gathered and the ambiguity set is updated, \( \tilde{V}^{(t)} \) approaches \( V^* \) from above.

VII. CONCLUSION

In this article, we presented a distributionally robust MPC strategy for Markov jump systems with unknown transition probabilities subject to general probabilistic constraints. We proved closed-loop constraint satisfaction, MSS, and consistency of the resulting controller for a broad range of data-driven ambiguity sets.

APPENDIX

A. Technical Lemma

**Lemma A.1 (Infimum convergence):** Consider a sequence of proper, lsc functions \( V^{(t)} : \mathbb{R}^n \to \mathbb{R}, t \in \mathbb{N} \), and a proper, lsc, level-bounded function \( V : \mathbb{R}^n \to \mathbb{R} \). Suppose that

i) (eventual upper bound) there exists a \( T \in \mathbb{N} \), such that for all \( t > T \), and for all \( u \), \( V^{(t)}(u) \geq V(u) \);

ii) (pointwise convergence) \( V^{(t)} \xrightarrow{\text{a.s.}} V \). That is, for all \( u \), \( \lim_{t \to \infty} V^{(t)}(u) = V(u) \).

Then, \( \lim_{t} \inf_{u} V^{(t)}(u) = \inf_{u} V(u) \).

**Proof:** By (i), it follows that for any sequence \( u_t \to \bar{u} \)

\[
\lim_{t} \inf_{u} V^{(t)}(u_t) = \lim_{t} \inf_{u} V^{(t)}(u) \geq \lim_{u \to \bar{u}} \inf_{u} V(u) \geq V(\bar{u})
\]

where the first inequality follows from Condition (i), and the second inequality follows from lower semicontinuity of \( V \). Moreover, fixing \( (u_t)_{t \in \mathbb{N}} \) to be the constant sequence \( u_t = \bar{u} \), it follows from (ii) that \( \lim_{t} \sup_{u} V^{(t)}(u_t) = \lim_{t} V^{(t)}(\bar{u}) \leq V(\bar{u}) \). Invoking [73, Prop. 7.2], we conclude that \( V^{(t)} \xrightarrow{\text{s.a.}} V \), i.e., \( V^{(t)} \) epiconverges to \( V \). Second, from Condition (i) and the level-boundedness of \( V \), it follows that \( (V^{(t)}(u))_{t \in \mathbb{N}} \) is eventually level bounded [73, Ex. 7.32]. The claim then follows from [73, Th. 7.33].
B. Deferred proofs

Proof of Lemma V.6.

Let \((z_t)_{t \in \mathbb{N}} = (x_t, s_t, \beta_t, w_t)_{t \in \mathbb{N}}\) denote the stochastic process satisfying dynamics (25), for some initial state \(z_0 \in \text{dom} \, V\). For ease of notation, let us define \(V_t := V(z_t), t \in \mathbb{N}\). Due to nonnegativity of \(V\)
\[
E \left[ \sum_{t=0}^{k-1} c \|x_t\|^2 \right] \leq E \left[ V_k + \sum_{t=0}^{k-1} c \|x_t\|^2 \right] \\
= E \left[ V_k - V_0 + \sum_{t=0}^{k-1} c \|x_t\|^2 \right] + V_0
\]
where the second equality follows from the fact that \(V_0\) is deterministic. By linearity of the expectation, we can in turn write
\[
E \left[ V_k - V_0 + \sum_{t=0}^{k-1} c \|x_t\|^2 \right] = E \left[ \sum_{t=0}^{k-1} V_{t+1} - V_t + c \|x_t\|^2 \right] \\
= \sum_{t=0}^{k-1} E \left[ V_{t+1} - V_t + c \|x_t\|^2 \right].
\]
Therefore,
\[
E \left[ c \sum_{t=0}^{k-1} \|x_t\|^2 \right] - V_0 \leq \sum_{t=0}^{k-1} E \left[ V_{t+1} - V_t + c \|x_t\|^2 \right] + c E \left[ \|x_k\|^2 \right].
\]
Recall that \(\beta_t\) denotes the coordinate of \(\beta\) corresponding to the risk measures in the cost function (20). Defining the event \(E_t := \{ \omega \in \Omega \mid P_{w_t}(\omega) \in A_{\beta_t}(s_t(\omega), w_t(\omega)) \}\), and its complement \(\bar{E}_t := \Omega \setminus E_t\), we can use the law of total expectation to write
\[
E \left[ V_{t+1} - V_t \right] = E \left[ V_{t+1} - V_t \mid E_t \right] P[E_t] \\
+ E \left[ V_{t+1} - V_t \mid \bar{E}_t \right] P[\bar{E}_t].
\]
By Condition (8), \(P[\bar{E}_t] < \beta_t\). From Conditions (i) and (ii), it follows that \(z_t \in \text{dom} \, V \forall t \in \mathbb{N} \cap [0,k]\) and that there exists a \(\bar{V} \geq 0\) such that \(V(z) \leq \bar{V}\), for all \(z \in \text{dom} \, V\). Therefore, \(E[\bar{V}_{t+1} - V_t] \leq \bar{V}\). Finally, by Condition (ii), \(E[\bar{V}_{t+1} - V_t] \leq \bar{V}\) \([\|x_t\|^2 \mid E_t]\). Thus,
\[
E \left[ V_{t+1} - V_t \right] \leq E \left[ -c \|x_t\|^2 \mid E_t \right] P[E_t] + \bar{V} \beta_t.
\]
This allows us to simplify expression (32) as
\[
E \left[ c \sum_{t=0}^{k-1} \|x_t\|^2 \right] - V_0 \leq \sum_{t=0}^{k-1} E \left[ -c \|x_t\|^2 \mid E_t \right] P[E_t] + \bar{V} \beta_t,
\]
which remains finite as \(k \to \infty\), since \((\beta_t)_{t \in \mathbb{N}}\) is summable. Thus, necessarily
\[
\lim_{t \to \infty} E \left[ \|x_t\|^2 \right] = 0.
\]

Proof of Lemma V.7.

First, note that using the monotonicity of coherent risk measures [34, Sec. 6.3, (R2)], a straightforward inductive argument allows us to show that under Condition (i)
\[
T \bar{V}_N \leq \tilde{V}_N \quad \forall N \in \mathbb{N}.
\]
Since \(\mathcal{Z} \subseteq \text{dom} \, \tilde{V}_N\), recall that by definition (24), we have for any \(z = (x, s, \beta, w) \in \mathcal{Z}\) that
\[
\tilde{V}_N(z) = \ell(x, \bar{r}_N(z), w) + \rho_{w,s}^\beta \left[ \tilde{V}_{N-1} \left( \bar{f}^{\bar{r}_N}(z, v), v \right) \right]
\]
where \(\beta\) denotes the component of \(\beta\) corresponding to the cost. Therefore, we may write
\[
\rho_{w,s}^\beta \left[ \tilde{V}_{N} \left( \bar{f}^{\bar{r}_N}(z, v), v \right) \right] = \rho_{w,s}^\beta \left[ \tilde{V}_{N-1} \left( \bar{f}^{\bar{r}_N}(z, v), v \right) \right] - \ell(x, \bar{r}_N(z), w) \\
\leq -\ell(x, \bar{r}_N(z), w) \leq -c \|x\|^2
\]
where the first inequality follows by (33) and monotonicity of coherent risk measures. The second inequality follows from Condition (ii). Combined with Condition (iii), this implies that \(V : z \mapsto V_N(z) + \delta_\mathcal{Z}(z)\) satisfies the conditions of Lemma V.6 and the assertion follows.

It will be convenient to define \(Q_N^{(t)}\) and \(Q_N\) as:
\[
Q_N^{(t)}(x, u, w) := \ell(x, u, w) + \rho_{w,s}^\beta [\tilde{V}_{N-1}(f(x, u, v), v)]
\]
\[
Q_N(x, u, w) := \ell(x, u, w) + E_{P_{w}}[\tilde{V}_{N-1}(f(x, u, v), v)|x, w]
\]
and let \(\tilde{U}^{(t)}(x, z) = \tilde{U}(x, s_t, \beta_t, w)\), so we may write
\[
\tilde{V}_N^{(t)}(x, w) = \inf_{u \in \bar{U}^{(t)}(x, w)} Q_N^{(t)}(x, u, w)
\]
\[
V_N(x, w) = \inf_{u \in \bar{U}(x, w)} Q_N(x, u, w).
\]

Proof of Theorem V.9:

We will show (27) by induction on \(N\). For \(N = 0\), we have \(\tilde{V}_0^{(t)} \equiv V_t \equiv V_0\), thus, (27) holds with \(\gamma_0^{(t)} = 0\), \forall t \in \mathbb{N}\), and the claim holds trivially. For the induction step, we define the events
\[
A^{(t)} := \{ P_{w} \in \cap_{t=1}^{N} A_{\beta_t}(s_t, w) \forall w \in W \}
\]
\[
B_N^{(t)} := \tilde{V}_N^{(t)}(x, w) \geq V_N(x, w) \forall (x, w) \in \mathbb{R}^n_x \times W
\]
for \(N \in \mathbb{N}, t \in \mathbb{N}\). The induction hypothesis now reads
\[
P[B_N^{(t)}] \geq \gamma_N^{(t)} = d \sum_{k=0}^{N-1} \|\beta_{t+k}\|_1 \forall t \in \mathbb{N}
\]
along with Goal (i) that implies \([B_N^{(t)}] \geq \gamma_N^{(t)} \forall t \in \mathbb{N}\).

Given the occurrence of event \(B_{N-1}^{(t)}\), the monotonicity of risk measures [34, Sec. 6.3, (R2)] ensures that \(Q_N^{(t)}(x, u, w) \geq \ell(x, u, w) + \rho_{w,s}^\beta [\tilde{V}_{N-1}(f(x, u, v), v)]\) for all \((x, u, w) \in \mathbb{R}^n_x \times \mathbb{R}^n_w \times W\), and \(t \in \mathbb{N}\). Furthermore, conditional on event \(A^{(t)}\), (16) implies that \(\rho_{w,s}^\beta \geq E_{P_{w}}\) and \(\rho_{s_t,w}^{\beta_t} \), uniformly. Combining this fact with (5) and (19), we obtain the implication
\[
B_N^{(t+1)}, A^{(t)} \Rightarrow Q_N^{(t+1)}(x, u, w) \geq Q_N(x, u, w)
\]
\[
\tilde{U}^{(t+1)}(x, w) \leq U(x, w)
\]
∀(x, u, w) ∈ \mathbb{R}^n_x × \mathbb{R}^n_u × W, and hence,
\[ \hat{V}_N^{(t)}(x, w) = \min_{u \in \mathcal{U}(x, w)} Q_N^{(t)}(x, u, w) \geq \min_{u \in \mathcal{U}(x, w)} Q_N(x, u, w) = V_N(x, w) \]
which describes exactly the event $B_N^{(t)}$. Thus, we have shown that $\mathbb{P}[B_N^{(t)}] ≥ \mathbb{P}[A^{(t)}, B_N^{(t-1)}]$. By the union bound, we now obtain
\[ \mathbb{P}[B_N^{(t)}] ≥ \mathbb{P}[A^{(t)}, B_N^{(t-1)}] ≥ 1 - \left( \mathbb{P}[\neg A^{(t)}] + \mathbb{P}[\neg B_N^{(t-1)}] \right) \geq 1 - (d\|\beta_t\|_1 + \gamma_{N-1}^{(t-1)}) \]
where in the final inequality, $\mathbb{P}[\neg B_N^{(t-1)}]$ was bounded using the induction hypothesis (37) and $\mathbb{P}[\neg A^{(t)}]$ was replaced by another application of the union bound:
\[ \mathbb{P}[\neg A^{(t)}] = \mathbb{P}[\exists w \in W : P_{w'} \not\in \bigcap_{n=1}^{N_w} \mathcal{A}_{\beta_t}, (s_t, w)] \leq \sum_{w \in W} \sum_{n=1}^{N_w} \mathbb{P}[P_{w'} \not\in \mathcal{A}_{\beta_t}, (s_t, w)] \leq d \sum_{n=1}^{N_w} \beta_t \leq d\|\beta_t\|_1. \]}

Thus, substituting the expression for $\gamma_{N-1}^{(t-1)}$ from the induction hypothesis (37) into the result (38), we obtain that (27) holds with
\[ \gamma_N := d\|\beta_t\|_1 + \gamma_{N-1}^{(t-1)} = d\sum_{k=1}^{N_w} \|\beta_t\|_1 \]}
which establishes (i). Under the conditions of (ii), namely that (28) holds, it follows from definition (36a) that:
\[ \mathbb{P}[\neg A^{(t)}] = \mathbb{P}[P_{w'} \not\in \mathcal{A}_{\beta_t}, (s_t, w)] \leq d\|\beta_t\|_1. \]

with $\beta_t^* := \max_{i \in \mathcal{I} \cap \mathcal{A}_{\beta_t}} \|\beta_t\|_1$. Statement (iii) is then established by the same inductive argument, replacing the expression for $\gamma_{N-1}^{(t-1)}$ in (37), and replacing (39) with (40).

**Proof of Theorem V.12.**

By Assumption V.8, we have for $N = 0$ that $\hat{V}_N^{(t)} \equiv V_0 \equiv \hat{V}_t$ and there is nothing to prove. The general case, $N > 0$, is proved by induction. Assume that (30) holds for some $N ≥ 0$. We will now demonstrate that this implies that it also holds for $N + 1$. To this end, we will show that the sequence $(Q_N^{(t)}(x, \cdot, w))_{w \in \mathcal{I}}$ and the function $Q_{N+1}^{(t)}(x, \cdot, w)$, satisfy the conditions of Lemma A.1. Under Assumption II.3, and using [73, Th. 3.31], it follows from [32, Prop. 2] that $Q_N$ and $Q_{N+1}$ are proper, lsc, and level-bounded in $u$ locally uniformly in $x$, for all $u \in W$. Let us introduce the shorthand for the worst-case conditional distribution
\[ p_t^*(w) = \arg \max_{P \in \mathcal{A}_{\beta_t}} \sum_{w \in W} p_w \hat{V}_N^{(t+1)}(f(x, u, v), v) \]
where we have omitted the dependence on the constant $x$ and $w$. Combining Corollary V.10 with Assumption III.5, the Borel–Cantelli lemma [46, Th. 4.3] guarantees that w.p. 1, there exists a finite $T_N \in \mathcal{I}$, such that for all $t > T_N$, $P_{w'} \in \mathcal{A}_{\beta_t}(s_t, w)$, for all $w \in W$ and $t \in [1, N_w]$, and furthermore, $\hat{V}_N^{(t)} ≥ V_N$, which implies that $Q_{N+1}^{(t)} ≥ Q_{N+1}$. Moreover, by the induction hypothesis [i.e., (30) holds for $N$], there exists for every $ε > 0$, a $T_t ≥ T_N$, such that for all $t > T_t$
\[ Q_N^{(t)}(x, u, w) - Q_{N+1}(x, u, w) \leq \sum_{w \in W} p_t^*(w)(\hat{V}_N^{(t+1)}(f(x, u, v), v) - P_{w'}V_N(f(x, u, v), v)) \]
where the final inequality is due to Assumption V.11 and the fact that for all $t > T_N$, $P_{w'} \in \mathcal{A}_{\beta_t}(s_t, w)$. As $t \to 0$, the first term in (41) can be made arbitrarily small by increasing $t$, provided that $V_N(f(x, u, v), v) < \infty$, for all $v \in W$, hence establishing pointwise convergence $\lim_{t \to 0} Q_N^{(t)} \to Q_{N+1}$ whenever dom $V_N$ is RCI for (1), which in turn holds if $\mathcal{A}_t$ is RCI by Proposition V.4. The sequence $(Q_N^{(t)}(x, \cdot, w))_{w \in \mathcal{I}}$ and the function $Q_{N+1}(x, \cdot, w)$, thus, satisfy the conditions of Lemma A.1, which establishes (30) for $N + 1$.

**References**

[1] M. Schuurmans and P. Patrinos, “Learning-based distributionally robust model predictive control of Markovian switching systems with guaranteed stability and recursive feasibility,” Sep. 2020, arXiv:2009.04422.

[2] B. Kouvaritakis and M. Cannon, Model Predictive Control Advanced Textbooks in Control and Signal Processing, Cham, Switzerland: Springer, 2016.

[3] A. Mesbah, “Stochastic model predictive control: An overview and perspective for future research,” IEEE Control Syst. Mag., vol. 36, no. 6, pp. 30–44, Dec. 2016.

[4] J. B. Rawlings, D. Q. Mayne, and M. M. Diehl, Model Predictive Control: Theory, Computation, and Design, 2nd ed. Madison, WI, USA: Nob Hill, 2017.

[5] P. M. 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.

[6] J. Dupačová, “The minimax approach to stochastic programming and an illustrative application,” Stochastics, vol. 20, pp. 73–88, Jan. 1987.

[7] B. P. C. Van Parys, P. M. Esfahani, and D. Kuhn, “From data to decisions: Distributionally robust optimization is optimal,” Manage. Sci., vol. 67, no. 6, pp. 3387–3402, 2021.

[8] R. Gao and A. Kleywegt, “Distributionally robust stochastic optimization with Wasserstein distance,” Math. Operations Res., Aug. 2022, doi: 10.1287/moor.2022.1275.

[9] W. Wiesemann, D. Kuhn, and M. Sim, “Distributionally robust convex optimization,” Operations Res., vol. 62, pp. 1538–1576, Dec. 2014.

[10] D. Bertsimas, V. Gupta, and N. Kallus, “Data-driven robust optimization,” Math. Program., vol. 167, pp. 235–292, Feb. 2018.

[11] M. Schuurmans, P. Sopasakis, and P. Patrinos, “Safe learning-based control of stochastic jump linear systems: A distributionally robust approach,” in Proc. 58th IEEE Conf. Decis. Control, Dec. 2019, pp. 6498–6503.

[12] P. Coppens, M. Schuurmans, and P. Patrinos, “Data-driven distributionally robust LQR with multiplicative noise,” in Proc. Learn. Dyn. Control, Jul. 2020, pp. 521–530.

[13] I. Yang, “Wasserstein distributionally robust stochastic control: A data-driven approach,” IEEE Trans. Autom. Control, vol. 66, no. 8, pp. 3863–3870, Aug. 2021.

[14] A. Hakobyan and I. Yang, “Wasserstein distributionally robust motion control for collision avoidance using conditional value-at-risk,” IEEE Trans. Robot., vol. 38, no. 2, pp. 939–957, Apr. 2022.

[15] A. Hakobyan and I. Yang, “Distributionally robust risk map for learning-based motion planning and control: A semidefinite programming approach,” IEEE Trans. Robot., pp. 1–20, 2022, doi: 10.1109/TRO.2022.3200156.
[67] M. V. Kothare, V. Balakrishnan, and M. Morari, “Robust constrained model predictive control using linear matrix inequalities,” *Automatica*, vol. 32, no. 10, pp. 1361–1379, 1996.

[68] MOSEK ApS, “MOSEK optimization toolbox for MATLAB manual. Version 8.1.,” 2017. Accessed: Nov. 15, 2021. [Online]. Available: http://docs.mosek.com/8.1/toolbox/index.html

[69] S. Diamond and S. Boyd, “CVXPY: A Python-embedded modeling language for convex optimization,” *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 2909–2913, 2016.

[70] A. Shapiro, “On Duality Theory of Conic Linear Problems,” in *Semi-Infinite Programming*, P. Pardalos, M. A. Goberna, and M. A. López, Eds., Boston, MA, USA: Springer, 2001, pp. 135–165.

[71] G. C. Pflug and A. Pichler, *Multistage Stochastic Optimization*. Cham, Switzerland: Springer, 2014.

[72] M. Schuurmans, A. Katriniok, H. E. Tseng, and P. Patrinos, “Learning-based risk-averse model predictive control for adaptive cruise control with stochastic driver models,” in *Proc. IFAC World Congr.*, 2020, pp. 15337–15342.

[73] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, vol. 317 of *Grundlehren Der Mathematischen Wissenschaften*. Berlin, Germany: Springer, 1998.

[74] M. Schuurmans and P. Patrinos, “A general framework for learning-based distributionally robust MPC of Markov jump systems,” Jan. 2023, arXiv: 2106.00561, doi: 10.48550/arXiv.2106.00561.

Mathijs Schuurmans received the B.Sc. degree in electrical and mechanical engineering and the M.Sc. degree in mathematical engineering in 2016 and 2018, respectively, from KU Leuven, Leuven, Belgium, where he is currently working toward the Ph.D. degree in electrical engineering with the Department of Electrical Engineering (ESAT).

His research focuses on data-driven model predictive control of stochastic systems, focusing on distributionally robust approaches for safety-critical applications in autonomous driving.

Panagiotis (Panos) Patrinos received the M.Eng. degree in chemical engineering, M.S. degree in applied mathematics, and the Ph.D. degree in control and optimization from the National Technical University of Athens, Athens, Greece, in 2003, 2005, and 2010, respectively. He is currently an Associate Professor with the Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium. In 2014, he was a Visiting Professor with Stanford University, Stanford, CA, USA. After his Ph.D., he was a Postdoc with the University of Trento, Trento, Italy, and IMT Lucca, Lucca, Italy, where he became an Assistant Professor in 2012. His current research interests include the intersection of optimization, control and learning, theory and algorithms for structured nonconvex optimization, and learning-based, model predictive control with a wide range of applications including autonomous vehicles, machine learning, and signal processing.

Dr. Patrinos was the corecipient of the 2020 Best Paper Award in the *International Journal of Circuit Theory and Applications*. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.