2020 59th IEEE Conference on Decision and Control (CDC)
Jeju Island, Republic of Korea, December 14-18, 2020

Learning-Based Distributionally Robust Model Predictive Control of Markovian Switching Systems with Guaranteed Stability and Recursive Feasibility

Mathijs Schuurmans and Panagiotis Patrinos

***Abstract*— We present a data-driven model predictive control scheme for chance-constrained Markovian switching 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.**

I. Introduction

Stochasticity is inherent to processes emerging in a multitude of applications. Nevertheless, we are often required to performantly control such systems, using only the limited information that is available. Over the past decades, the decreasing cost of data acquisition, transmission and storage has caused a surge in research interest in data-driven approaches towards control. More recently, as the focus in research is gradually shifting towards real-life, safety-critical applications, there has been an increasing concern for safety guarantees of such data-driven methods, which are valid in a finite-data regime (see [1] for a recent survey).

In this paper, we focus on safe, learning-based control of Markovian switching systems. Control of this class of systems is widely studied and has been used to model systems stemming from a wide range of applications [2]–[4].

Many of these approaches, however, require knowledge of the transition kernel governing the switching behavior of these systems, although in practice, such information is typically unavailable. Although recently, data-driven methods have been proposed to address this problem [5], [6], relatively little attention has gone to providing *a priori* guarantees on stability and constraint satisfaction.

Leveraging the framework of risk-averse model predictive control (MPC) [7]–[9] and its connection to distributionally robust (DR) optimization [10], we propose a learning-based DR MPC scheme, which is provably stable, recursively feasible and uses data gathered during operation to improve performance, rendering it less conservative than more traditional robust approaches [11]. Despite its growing popularity [9], [12]–[15], this approach has remained relatively unexplored for chance-constrained stochastic optimal control problems with ambiguity in the estimation of conditional distributions, in particular for Markovian switching dynamics.

We summarize our contributions as follows. (i) We propose a general data-driven, DR MPC scheme for Markov switching systems with unknown transition probabilities, which is compatible with the recently developed framework of risk-averse MPC [7], [8]. 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. (ii) We state the problem in terms of an augmented state vector of constant dimension, which summarizes the available information at every time. This allows us to formulate the otherwise time-varying optimal control problem as a dynamic programming recursion. (iii) We provide sufficient conditions for recursive feasibility and mean-square stability of the proposed MPC scheme, with respect to the true-but-unknown probability distributions.

A. Notation

Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{N}_{>0} := \mathbb{N} \setminus 0$. For two naturals $a, b \in \mathbb{N}, a \leq b$, we denote $\mathbb{N}_{[a,b]} := \{n \in \mathbb{N} \mid a \leq n \leq b\}$ and similarly, we introduce the shorthand $w_{[a,b]} := (w_t)_{t=a}^{b}$ to denote a sequence of variables. We denote the extended real line by $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ and the set of *nonnegative* (extended) real numbers by $\mathbb{R}_+$ (and $\overline{\mathbb{R}}_+$). 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. 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 whenever confusion with time indices is possible. $\mathbf{vec}(M)$ denotes the vertical concatenation of the columns of a matrix M. We denote the vector in $\mathbb{R}^k$ with all elements one as $\mathbf{1}_k := (1)_{i=1}^{k}$ and the probability simplex of dimension k as $\Delta_k := \{p \in \mathbb{R}_+^k \mid p^\top \mathbf{1}_k = 1\}$. We define the indicator function as $1_{x=y} = 1$ if $x = y$ and 0 otherwise. Similarly, the characteristic function $\delta_{\mathcal{X}} : \mathbb{R}^n \to \overline{\mathbb{R}}$ of a set $\mathcal{X} \subseteq \mathbb{R}^n$ is defined by $\delta_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and ∞ otherwise.

M. Schuurmans and P. Patrinos are with the Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. Email: {mathijs.schuurmans, panos.patrinos}@esat.kuleuven.be

This work was supported by: FWO projects: No. G086318N; No. G086518N; Fonds de la Recherche Scientifique – FNRS, the Fonds Wetenschappelijk Onderzoek–Vlaanderen under EOS Project No. 30468160 (SeLMA), Research Council KU Leuven C1 project No. C14/18/068 and the Ford–KU Leuven Research Alliance project No. KUL0023.

II. Problem statement

We consider discrete time Markovian switching systems with dynamics of the form

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

where $x_t \in \mathbb{R}^{n_x}, u_t \in \mathbb{R}^{n_u}$ are the state and control action at time t, respectively, and $w_{t+1} : \Omega \to W$ is a random variable drawn from a discrete-time, time-homogeneous Markov chain $\mathbf{w} := (w_t)_{t\in\mathbb{N}}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values on $W := \mathbb{N}_{[1,M]}$. We refer to w_t as the *mode* of the Markov chain at time t. The *transition kernel* governing the Markov chain is denoted by $P = (P_{ij})_{i,j\in W}$, where $P_{ij} = \mathbb{P}[w_t = j \mid w_{t-1} = i]$. We assume that the state x_t and mode w_t are observable at time t. For a given state-mode pair $(x, w) \in \mathbb{R}^{n_x} \times W$, we constrain the control action u to the set $\mathcal{U}(x, w)$, defined as

$$\begin{aligned}\mathcal{U}(x,w) := \{u \in \mathbb{R}^{n_u} :\\ \mathbb{P}[g_i(x,u,w,v) > 0 \mid x, w] \le \alpha_i, \forall i \in \mathbb{N}_{[1,n_g]}\},\end{aligned} \tag{2}$$

where $v \sim P_w$ is randomly drawn from the Markov chain $\mathbf{w}$ in mode w, and $g_i : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times W^2 \to \mathbb{R}$ are constraint functions with corresponding constraint violation rates α_i. By appropriate choices of α_i and g_i, constraint (2) can be used to encode robust constraints ($\alpha_i = 0$) or chance constraints ($0 < \alpha_i < 1$) on the state, the control action, or both. Note that the formulation (2) additionally covers chance constraints on the successor state $f(x, u, v)$ under input u, conditioned on the current values x and w.

Ideally, our goal is to synthesize – by means of a stochastic MPC scheme – a stabilizing control law $\kappa_N : \mathbb{R}^{n_x} \times W \to \mathbb{R}^{n_u}$, 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 (a.s.) 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 $\pi = (\pi_k)_{k=0}^{N-1}$, referred to as a *policy* of length N. Given a stage cost $\ell : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times W \to \mathbb{R}_+$, and a terminal cost $V_\mathrm{f} : \mathbb{R}^{n_x} \times W \to \mathbb{R}_+$ and corresponding terminal set $\mathcal{X}_\mathrm{f}$: $\bar{V}_\mathrm{f}(x, w) := V_\mathrm{f}(x, w) + \delta_{\mathcal{X}_\mathrm{f}}(x)$, we can assign to each such policy π, a cost

$$V_N^\pi(x,w) := \mathbb{E}\big[\textstyle\sum_{k=0}^{N-1} \ell(x_k, u_k, w_k) + \bar{V}_\mathrm{f}(x_N, w_N)\big], \tag{3}$$

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

Definition II.1 (Stochastic OCP)**.** *For a given state-mode pair (x, w), the optimal cost of the stochastic OCP is*

$$V_N(x,w) = \min_\pi V_N^\pi(x,w) \tag{4a}$$

subject to

$$x_0 = x, w_0 = w, \pi = (\pi_k)_{k=0}^{N-1}, \tag{4b}$$

$$x_{k+1} = f(x_k, \pi_k(x_k, w_k), w_{k+1}), \tag{4c}$$

$$\pi_k(x_k, w_k) \in \mathcal{U}(x_k, w_k),\ \forall k \in \mathbb{N}_{[0,N-1]}. \tag{4d}$$

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

Let $(\pi_k^\star(x,w))_{k=0}^{N-1} \in \Pi_N(x, w)$, so that the stochastic MPC control law is given by $\kappa_N(x, w) = \pi_0^\star(x, w)$. Sufficient conditions on the terminal cost $\bar{V}_\mathrm{f}$ and its effective domain $\mathbf{dom}\, \bar{V}_\mathrm{f} = \mathcal{X}_\mathrm{f}$ to ensure mean-square stability of the closed-loop system, have been studied for a similar problem set-up in [3], 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 reality, P is typically not known but rather estimated from a finitely-sized sequence $w_{[0,t]}$ of observed values. Therefore, they are subject to some level of misestimation, commonly referred to as *ambiguity*. The goal of the proposed MPC scheme is to model this ambiguity explicitly, in order to be resilient against it, without disregarding available statistical information. To do so, we introduce the notion of a *learner state*, which is very similar in spirit to the concept of a *belief state*, commonly used in control of partially observed Markov decision processes [16]. It can be regarded as an internal state of the controller that stores all the information required to build a set of possible conditional distributions over the next state, given the observed data. We formalize this in the following assumption.

Assumption II.2 (Learning system)**.** *Given a sequence $w_{[0,t]}$ sampled from the Markov chain $\mathbf{w}$, we can compute (i) a statistic $s_t : W^{t+1} \to \mathcal{S} \subseteq \mathbb{R}^{n_s}$, accompanied by a vector of confidence parameters $\beta_t \in \mathcal{I} := [0,1]^{n_\beta}$, which admit recursive update rules $s_{t+1} = \mathcal{L}(s_t, w_t, w_{t+1})$ and $\beta_{t+1} = C(\beta_t)$, $t \in \mathbb{N}$; and (ii) an* ambiguity set *$\mathcal{A} : \mathcal{S} \times W \times [0,1] \rightrightarrows \Delta_M : (s, w, \beta) \mapsto \mathcal{A}_\beta(s, w)$, mapping s_t, w_t and an element β_t^i to a convex subset of the M-dimensional probability simplex Δ_M such that for all $t \in \mathbb{N}$,*

$$\mathbb{P}[P_{w_t} \in \mathcal{A}_{\beta_t^i}(s_t, w_t)] \ge 1 - \beta_t^i. \tag{5}$$

We will refer to s_t and β_t as the state of the learner and the confidence vector at time t, respectively.

Remark II.3 (confidence levels). Two points of clarification are in order. First, 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 n_g individual chance constraints as well as the cost function of the data-driven OCP (defined in Definition III.3 below). Accordingly, we will assume that $n_\beta = n_g + 1$.

Second, the confidence levels are completely exogenous to the system dynamics and can in principle be chosen to be any time-varying sequence satisfying the technical conditions discussed further (see Proposition III.1 and Assumption II.4). The requirement that the sequence $(\beta_t)_{t\in\mathbb{N}}$ can be written as the trajectory of a time-invariant dynamical system serves to facilitate theoretical analysis of the proposed scheme through dynamic programming. Meanwhile, it covers a large class of sequences one may reasonably choose, as illustrated in Example II.5.

We will additionally invoke the following assumption on the confidence levels when appropriate.

Assumption II.4. *The confidence dynamics $\beta_{t+1} = C(\beta_t)$ is chosen such that $\sum_{t=0}^\infty \beta_t^i < \infty, \forall i \in \mathbb{N}_{[1,n_\beta]}$.*

In other words, we will assume that the probability of obtaining an ambiguity set that contains the true conditional distribution (expressed by (5)) increases sufficiently fast. This assumption will be of crucial importance in showing stability (see Section IV-C).

To fix ideas, consider the following example of a learning system satisfying the requirements of Assumption II.2.

Example II.5 (Transition counts and ℓ_1-ambiguity). A natural choice for the learner state is to take $s_t = \mathbf{vec}(m(t))$, where $m(t) = (m_{wv}(t))_{w,v\in W} \in \mathbb{N}^{M\times M}$ contains the mode transition counts at time t. That is, $m_{wv}(t) = |\{\tau \in \mathbb{N}_{[1,t]} : w_{\tau-1} = w, w_\tau = v\}|$, for all $w, v \in W$. It is clear that we can indeed write $s_{t+1} = \mathcal{L}(s_t, w_t, w_{t+1})$. Furthermore, following [17], we can uniquely obtain ambiguity sets parametrized as

$$\mathcal{A}^{\ell_1}_{\beta_t^i}(s_t, w_t) := \{p \in \Delta_M : \|p - \hat{p}_{w_t}\|_1 \le r_{w_t}(s_t, \beta_t^i)\},$$

for $i \in \mathbb{N}_{[1,n_\beta]}$, where $\hat{p}_{w_t}$ is the empirical estimate of the w_t'th row of the transition kernel (initialized to the uniform distribution if no transitions originating in mode w_t have been observed) and the radii $r_{w_t}(s_t, \beta_t^i)$ are chosen to satisfy (5) by means of basic concentration inequalities, yielding a closed-loop expression $\mathcal{O}(\sqrt{-m_{w_t}(t)^{-1}\ln(\beta_t)})$, with $m_{w_t} = \sum_{v\in W} m_{w_t v}$. Finally, the sequence of the confidence levels β_t remains to be selected. One particular family of sequences satisfying the additional requirement of Assumption II.4 is $\beta_t = b(1+t)^{-q}$, with parameters $0 \le b \le 1$, $q > 1$. This sequence can be described by the recursion $\beta_{t+1} = b\beta_t(\beta_t^{1/q} + b^{1/q})^{-q}$, $\beta_0 = b$, satisfying the requirements of Assumption II.2. Additionally, it allows the radius of the ambiguity set to converge to zero with the corresponding sample size. △

Equipped with a generic learning system of this form, our aim is to find a data-driven approximation to the stochastic OCP defined by (4), which asymptotically attains the optimal cost while preserving stability and constraint satisfaction during closed-loop operation.

III. Data-driven model predictive control

Given a learning system satisfying Assumption II.2, we define the augmented state $z_t = (x_t, s_t, \beta_t, w_t) \in \mathcal{Z} := \mathbb{R}^{n_x} \times \mathcal{S}\times\mathcal{I}\times W$, which evolves over time according to the dynamics

$$z_{t+1} = \tilde{f}(z_t, u_t, w_{t+1}) := \begin{bmatrix} f(x_t,u_t,w_{t+1}) \\ \mathcal{L}(s_t,w_t,w_{t+1}) \\ C(\beta_t) \\ w_{t+1} \end{bmatrix}, \tag{6}$$

where $w_{t+1} \sim P_{w_t}$, for $t \in \mathbb{N}$. Consequently, our scheme will result in a feedback law $\kappa : \mathcal{Z} \to \mathbb{R}^{n_u}$. To this end, we will formulate a DR counterpart to the stochastic OCP (4), 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 (3), we first introduce an ambiguous conditional expectation operator, leading to a formulation akin to the Markovian risk measures utilized in [7], [18]. Consider a function $\xi : \mathcal{Z} \times W \to \overline{\mathbb{R}}$, defining a stochastic process $(\xi_t)_{t\in\mathbb{N}} = (\xi(z_t, w_{t+1}))_{t\in\mathbb{N}}$ on $(\Omega, \mathcal{F}, \mathbb{P})$, and suppose that the augmented state $z_t = z = (x, s, \beta, w)$ is given. For simplicity, let us assume for the moment that $\beta \in [0, 1]$ is scalar; the following definition can be repeated for each component in the general case. The ambiguous conditional expectation of $\xi(z, v)$, given z is

$$\begin{aligned}\rho^\beta_{s,w}[\xi(z,v)] &:= \max_{p\in\mathcal{A}_\beta(s,w)} \mathbb{E}_p[\xi(z,v)|z] \\ &= \max_{p\in\mathcal{A}_\beta(s,w)} \textstyle\sum_{v\in W} p_v \xi(z,v).\end{aligned} \tag{7}$$

Trivially, it holds that if the w'th row of the transition matrix lies in the corresponding ambiguity set: $P_w \in \mathcal{A}_\beta(s, w)$, then $\rho^\beta_{s,w}[\xi(z,v)] \ge \mathbb{E}_{P_w}[\xi(z,v)|z] = \sum_{v\in W} P_{wv}\xi(z,v)$. Note that the function $\rho^\beta_{s,w}$ defines a coherent risk measure [19, Sec. 6.3]. We say that $\rho^\beta_{s,w}$ is the risk measure *induced by* the ambiguity set $\mathcal{A}_\beta(s, w)$.

A similar construction can be carried out for the chance constraints (2). In their standard form, chance constraints lead to nonconvex, nonsmooth constraints. For this reason, they are commonly approximated using risk measures [20]. Particularly, the (conditional) *average value-at-risk* (at level $\hat{\alpha} \in (0, 1]$ and with reference distribution $p \in \Delta_M$) of ξ is the coherent risk measure

$$\begin{aligned}&\mathrm{AV@R}^p_{\hat{\alpha}}[\xi(z,v) \mid z] \\ &= \begin{cases} \min\limits_{t\in\mathbb{R}} t + 1/\hat{\alpha}\mathbb{E}_p\left[\xi(z,v) - t \mid z\right]_+, & \hat{\alpha} \ne 0 \\ \max_{v\in W}[\xi(z,v)], & \hat{\alpha} = 0. \end{cases}\end{aligned} \tag{8}$$

It can be shown (e.g., [19, sec. 6.2.4]) that if $p = P_w$, then the following implication holds tightly

$$\mathrm{AV@R}^p_{\hat{\alpha}}[\xi(z,v)|z] \le 0 \Rightarrow \mathbb{P}[\xi(z,v) \le 0|z] \ge 1 - \hat{\alpha}. \tag{9}$$

By exploiting the dual risk representation [19, Thm 6.5], the left-hand inequality in (9) can be formulated in terms of only linear constraints [8]. As such, it can be used as a tractable surrogate for the original chance constraints, given perfect probabilistic information. Accounting also for the ambiguity in the knowledge of P_w through $\mathcal{A}_\beta(s, w)$, we define

$$\bar{\rho}^{\beta,\hat{\alpha}}_{s,w}[\xi(z,v)] := \max_{p\in\mathcal{A}_\beta(s,w)} \mathrm{AV@R}^p_{\hat{\alpha}}[\xi(z,v) \mid z] \le 0. \tag{10}$$

The function $\bar{\rho}^{\beta,\hat{\alpha}}_{s,w}$ in turn defines a coherent risk measure.

We now present a condition on the choice of $\hat{\alpha}$ under which a constraint of the form (10) can be used as a tractable and safe approximation of a chance constraint when there is ambiguity in the probability distribution.

Proposition III.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 II.2, i.e., $\mathbb{P}[P_w \in \mathcal{A}_\beta(s, w)] \ge 1-\beta$. If the parameter $\hat{\alpha}$ is chosen to satisfy $0 \le \hat{\alpha} \le \frac{\alpha-\beta}{1-\beta} \le 1$, then, for an arbitrary function $g : \mathcal{Z} \times W \to \overline{\mathbb{R}}$, the following implication holds:*

$$\bar{\rho}^{\beta,\hat{\alpha}}_{s,w}[g(z,v)] \le 0, \text{ a.s.} \Rightarrow \mathbb{P}[g(z,v) \le 0|x,w] \ge 1-\alpha. \tag{11}$$

Proof. If $\bar{\rho}^{\beta,\hat{\alpha}}_{s,w}[g(z,v)] \le 0$, a.s., then by (9)-(10),

$$\mathbb{P}[g(z,v) \le 0 \mid x, w, P_w \in \mathcal{A}_\beta(s,w)] \ge 1 - \hat{\alpha}, \text{a.s.}$$

Therefore,

$$\begin{aligned}&\mathbb{P}[g(z,v) \le 0 \mid x, w] \\ &\ge \mathbb{P}[g(z,v) \le 0 \mid x, w, P_w \in \mathcal{A}_\beta(s,w)]\mathbb{P}[P_w \in \mathcal{A}_\beta(s,w)] \\ &\ge (1-\hat{\alpha})(1-\beta).\end{aligned}$$

Requiring that $(1-\hat{\alpha})(1-\beta) \geq (1-\alpha)$ then immediately yields the sought condition. □

Notice that the implication (11) in Proposition III.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 is revealed. Indeed, for a *given* learner state s and mode w, the ambiguity set $\mathcal{A}_\beta(s,w)$ is fixed and therefore, the outcome of the event $E = \{P_w \in \mathcal{A}_\beta(s,w)\}$ is determined. Whether (11) then holds for these fixed values depends on the outcome of E. This is naturally reflected through the above condition on $\hat{\alpha}$ which implies that $\hat{\alpha} \leq \alpha$, and thus tightens the chance constraints that are imposed for a *fixed* s. Hence, the possibility that for this particular s, the ambiguity set may not include the conditional distribution, is accounted for. This effect can be mitigated by decreasing β, at the cost of a larger ambiguity set. A more detailed study of this trade-off is left for future work.

B. Risk-averse optimal control

We are now ready to describe the DR counterpart to the OCP (4), which, when solved in receding horizon fashion, yields the proposed data-driven MPC scheme.

For a given augmented state $z = (x,s,\beta,w) \in \mathcal{Z}$, we use (10) to define the DR set of feasible inputs $\hat{\mathcal{U}}(z)$ in correspondence to (2). Without loss of generality, let us assume that the ith entry β^i in the confidence vector β corresponds to the ith constraint, so that

$$\hat{\mathcal{U}}(z)=\{u \mid \bar{\rho}^{\beta^i,\hat{\alpha}_i}_{s,w}[g_i(x,u,w,v)] \leq 0, \forall i \in \mathbb{N}_{[1,n_g]}\}. \quad (12)$$

Remark III.2. The parameters $\hat{\alpha}_i$ remain to be chosen in relation to the confidence levels β and the original violation rates α_i. In light of Proposition III.1, $\hat{\alpha}_i = \frac{\alpha_i - \beta^i}{1-\beta^i}$ yields the least conservative choice. This choice is valid as long as it is ensured that $\beta^i \leq \alpha_i$.

Let us denote the remaining element of the confidence vector β corresponding to the cost by β^0. Using (7), we can then express the DR cost of a policy $\pi = (\pi_k)_{k=0}^{N-1}$ as

$$\begin{aligned}\hat{V}^\pi_N(z) := \ell(x_0,u_0,w_0) + \rho^{\beta^0_0}_{s_0,w_0}\big[\ell(x_1,u_1,w_1) \\ + \rho^{\beta^0_1}_{s_1,w_1}\big[\cdots + \rho^{\beta^0_{N-2}}_{s_{N-2},w_{N-2}}\big[\ell(x_{N-1},u_{N-1},w_{N-1}) \\ + \rho^{\beta^0_N}_{s_N,w_N}[\hat{V}_\mathrm{f}(x_N,s_N,\beta_N,w_N)]\big]\cdots\big]\big], \quad (13)\end{aligned}$$

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

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

$$\hat{V}_N(z) = \min_\pi \hat{V}^\pi_N(z) \quad (14a)$$

subject to

$$(x_0,s_0,\beta_0,w_0) = z,\ \pi = (\pi_k)_{k=0}^{N-1}, \quad (14b)$$

$$z_{k+1} = \tilde{f}(z_k,\pi_k(z_k),w_{k+1}), \quad (14c)$$

$$\pi_k(z_k) \in \hat{\mathcal{U}}(z_k),\ \forall w_{[0,k]} \in W^k, \quad (14d)$$

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

We thus define the data-driven MPC law analogously to the stochastic case as

$$\hat{\kappa}_N(z) = \hat{\pi}^\star_0(z), \quad (15)$$

where $(\hat{\pi}^\star_k(z))_{k=0}^{N-1} \in \hat{\Pi}_N(z)$. At every time t, the data-driven MPC scheme thus consists of (i) computing a control action $u_t = \hat{\kappa}_N(z_t)$ and applying it to the system (1); (ii) observing the outcome of $w_{t+1} \in W$ and the corresponding next state $x_{t+1} = f(x_t,u_t,w_{t+1})$; and (iii) updating the learner state $s_{t+1} = \mathcal{L}(s_t,w_t,w_{t+1})$ and the confidence levels $\beta_{t+1} = C(\beta_t)$, gradually decreasing the size of the ambiguity sets.

Remark III.4 (Scenario tree representations). Since W is a finite set, the possible realizations of $w_{[0,N]}$, given z_0 can be enumerated such that the corresponding predicted states and controls can be represented on a scenario tree [21]. It therefore suffices to optimize over a finite number (equal to the number of nodes in the tree) of control actions instead of infinite-dimensional control laws. When represented in this manner, it is apparent that (14) falls within the class of *risk-averse, risk-constrained optimal control problems*, described in [8]. In particular, the constraints (14d) at stage k can be represented in the framework of [8] as *nested risk constraints* which are compositions of a set of conditional risk mappings. In this case, the composition consists of $k-1$ **max** operators over values on the nodes in the first stages and a conditional risk mapping based on (10) at stage k. This is in line with the observations of [22, Sec. 7.1]. Consequently, if the risk measures employed in the definition of the DR-OCP (14) belong to the broad family of conic risk measures and the dynamics $f(\cdot,\cdot,w)$, $w \in W$ are linear, then, using the reformulations in [8], (14) can be cast as a convex conic optimization problem. This is the case for many commonly used coherent risk measures, including the risk measure induced by the ℓ_1-ambiguity set discussed in Example II.5 (see [23] for a numerical case study). For nonlinear dynamics, the problem is no longer convex but can in practice still be solved effectively with standard NLP solvers.

IV. Theoretical Analysis

A. Dynamic programming

To facilitate theoretical analysis of the proposed MPC scheme, we represent (14) as a dynamic programming recursion, similarly to [7]. We define the Bellman operator $\mathbf{T}$ as $\mathbf{T}(\hat{V})(z) := \mathbf{min}_{u \in \hat{\mathcal{U}}(z)} \ell(x,u,w) + \rho^{\beta^0}_{s,w}[\hat{V}(\tilde{f}(z,u,v))]$, where $z = (x,s,\beta,w) \in \mathcal{Z}$ are fixed quantities and $v \sim P_w$. We denote by $\mathbf{S}(\hat{V})(z)$ the corresponding set of minimizers. The optimal cost $\hat{V}_N$ of (14) is obtained through the iteration,

$$\hat{V}_k = \mathbf{T}\,\hat{V}_{k-1},\ \hat{V}_0 = \hat{V}_\mathrm{f},\ k \in \mathbb{N}_{[1,N]}. \quad (16)$$

Similarly, $\hat{\mathcal{Z}}_k := \mathbf{dom}\, \hat{V}_k$ is given recursively by

$$\hat{\mathcal{Z}}_k = \left\{ z \,\middle|\, \exists u \in \hat{\mathcal{U}}(z) : \tilde{f}(z,u,v) \in \hat{\mathcal{Z}}_{k-1},\ \forall v \in W \right\}.$$

Now consider the stochastic closed-loop system

$$z_{t+1} = \tilde{f}^{\hat{\kappa}_N}(z_t, w_{t+1}) := \tilde{f}(z_t, \hat{\kappa}_N(z_t), w_{t+1}), \tag{17}$$

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

B. Constraint satisfaction and recursive feasibility

In order to show existence of $\hat{\kappa}_N \in \mathbf{S}\,\hat{V}_{N-1}$ at every time step, Proposition IV.2 will require that $\hat{\mathcal{X}}_\mathrm{f}$ is a robust control invariant set. We define robust control invariance for the augmented control system under consideration as follows.

Definition IV.1 (Robust invariance)**.** *A set $\mathcal{R} \subseteq \mathcal{Z}$ is a robust control invariant (RCI) set for the system (6) if for all $z \in \mathcal{R}$, $\exists u \in \hat{\mathcal{U}}(z)$ such that $\tilde{f}(z,u,v) \in \mathcal{R}, \forall v \in W$. Similarly, $\mathcal{R}$ is a robust positive invariant (RPI) set for the closed-loop system (17) if for all $z \in \mathcal{R}$, $\tilde{f}^{\hat{\kappa}_N}(z,v) \in \mathcal{R}$, $\forall v \in W$.*

Since $\hat{\mathcal{U}}$ consists of conditional risk constraints, Definition IV.1 provides a DR counterpart to the notion of *stochastic* robust invariance in [24]. It is thus less conservative than the more classical notion of robust invariance for a set $\mathcal{R}_x$, obtained by imposing that $x \in \mathcal{R}_x \Rightarrow \exists u : g_i(x,u,w,v) \leq 0$, for all $i \in \mathbb{N}_{[1,n_g]}, w, v \in W$. In fact, $\mathcal{R}_x \times \mathcal{S} \times \mathcal{I} \times W$ is covered by Definition IV.1.

Proposition IV.2 (Recursive feasibility)**.** *If $\hat{\mathcal{X}}_\mathrm{f}$ is an RCI set for (6), then (14) is recursively feasible. That is, feasibility of DR-OCP (14) for some $z \in \mathcal{Z}$, implies feasibility for $z^+ = \tilde{f}^{\hat{\kappa}_N}(z,v)$, for all $v \in W, N \in \mathbb{N}_{>0}$.*

Proposition IV.2 follows from a standard inductive argument. We refer to the technical report [25] for the proof.

Corollary IV.3 (Chance constraint satisfaction)**.** *If Proposition IV.2 holds, then by Proposition III.1, the stochastic process $(z_t)_{t\in\mathbb{N}} = (x_t, s_t, \beta_t, w_t)_{t\in\mathbb{N}}$ satisfying dynamics (17) satisfies the nominal chance constraints*

$$\mathbb{P}[g_i(x_t, \hat{\kappa}_N(z_t), w_{t+1}) > 0 \mid x_t, w_t] < \alpha_i,$$

a.s., for all $i \in \mathbb{N}_{[1,n_g]}, t \in \mathbb{N}$.

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 $\mathcal{A}_\beta(s,w) = \Delta_M$ for all s, w, β. It is apparent from (10) and (12), 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

We will now provide sufficient conditions on the control setup under which the origin is mean-square stable (MSS) for (17), i.e., $\mathbf{lim}_{t\to\infty} \mathbb{E}[\|x_t\|^2] = 0$ for all x_0 in some specified compact set containing the origin.

Our main stability result, stated in Theorem IV.5, hinges in large on the following lemma, which relates *risk-square stability* [7, Lem. 5] of the origin for the autonomous system (17) to stability in the mean-square sense, based on the statistical properties of the ambiguity sets.

Lemma IV.4 (DR MSS condition)**.** *Suppose that Assumption II.4 holds and that there exists a nonnegative, proper function $V : \mathcal{Z} \to \overline{\mathbb{R}}_+$, such that (i) $\mathbf{dom}\, V \subseteq \mathcal{X} \times \mathcal{S} \times \mathcal{I} \times W$ is RPI for system (17), where $\mathcal{X} \subset \mathbb{R}^{n_x}$ is a compact set containing the origin; (ii) $\rho^{\beta^0}_{s,w}[V(\tilde{f}^{\hat{\kappa}_N}(z,v),v)] - V(z) \leq -c\|x\|^2$, for some $c > 0$, for all $z \in \mathbf{dom}\, V$; (iii) V is uniformly bounded on its domain, i.e., there exists a $\bar{V} \geq 0$, such that $V(z) \leq \bar{V}$ for all $z \in \mathbf{dom}\, V$. Then, $\mathbf{lim}_{t\to\infty} \mathbb{E}[\|x_t\|^2] = 0$ for all $z_0 \in \mathbf{dom}\, V$, 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 (17).*

Proof. See Appendix. □

Theorem IV.5 (MPC stability)**.** *Suppose that Assumption II.4 is satisfied, and let $\bar{\mathcal{Z}} := \bar{\mathcal{X}} \times \mathcal{S} \times \mathcal{I} \times W \subseteq \mathbf{dom}\, \hat{V}_N$ be an RPI set for (17), where $\bar{\mathcal{X}} \subseteq \mathbb{R}^{n_x}$ is a compact set containing the origin. Assume furthermore that the following statements hold: (i) $\mathbf{T}\hat{V}_\mathrm{f} \leq \hat{V}_\mathrm{f}$; (ii) $c\|x\|^2 \leq \ell(x,u,w)$ for some $c > 0$ for all $z = (x,w,\beta,s) \in \mathbf{dom}\, \hat{V}_N$ and all $u \in \hat{\mathcal{U}}(z)$. (iii) $\hat{V}_N$ is locally bounded in $\bar{\mathcal{Z}}$. Then, the origin is MSS for the MPC-controlled system (17), for all initial states $z_0 \in \bar{\mathcal{Z}}$.*

Proof. The proof is along the lines of that of [7, thm. 6] and shows that $\hat{V}_N$ satisfies the conditions of Lemma IV.4. Details are in the Appendix. □

V. Conclusions

We presented a distributionally robust MPC strategy for Markovian switching systems with unknown transition probabilities subject to general chance constraints. using data-driven, high-confidence ambiguity sets, we derive a DR counterpart to a nominal stochastic MPC scheme. We show that the resulting scheme provides a priori guarantees on closed-loop constraint satisfaction and mean-square stability of the true system, without requiring explicit knowledge of the transition probabilities.

References

[1] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, "Learning-Based Model Predictive Control: Toward Safe Learning in Control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, 2020.

[2] O. L. d. V. Costa, M. D. Fragoso, and R. P. Marques, *Discrete-time Markov jump linear systems*. Probability and its applications, London: Springer, 2005.

[3] P. Patrinos, P. Sopasakis, H. Sarimveis, and A. Bemporad, "Stochastic model predictive control for constrained discrete-time Markovian switching systems," *Automatica*, vol. 50, pp. 2504–2514, Oct. 2014.

[4] D. Bernardini and A. Bemporad, "Scenario-based model predictive control of stochastic constrained linear systems," in *48th IEEE Conference on Decision and Control (CDC) Held Jointly with 2009 28th Chinese Control Conference*, pp. 6333–6338, IEEE, Dec. 2009.

[5] R. L. Beirigo, M. G. Todorov, and A. M. S. Barreto, "Online TD(λ) for discrete-time Markov jump linear systems," in *57th IEEE Conference on Decision and Control (CDC)*, pp. 2229–2234, Dec. 2018.

[6] S. He, M. Zhang, H. Fang, F. Liu, X. Luan, and Z. Ding, "Reinforcement learning and adaptive optimization of a class of Markov jump systems with completely unknown dynamic information," *Neural Computing and Applications*, Apr. 2019.

[7] P. Sopasakis, D. Herceg, A. Bemporad, and P. Patrinos, "Risk-averse model predictive control," *Automatica*, vol. 100, pp. 281–288, Feb. 2019.

[8] P. Sopasakis, M. Schuurmans, and P. Patrinos, "Risk-averse risk-constrained optimal control," in *18th European Control Conference (ECC)*, pp. 375–380, June 2019.

[9] S. Singh, Y.-L. Chow, A. Majumdar, and M. Pavone, "A Framework for Time-Consistent, Risk-Sensitive Model Predictive Control: Theory and Algorithms," *arXiv:1703.01029*, Apr. 2018.

[10] H. Rahimian and S. Mehrotra, "Distributionally Robust Optimization: A Review," *arXiv:1908.05659*, Aug. 2019.

[11] A. Ben-Tal, L. El Ghaoui, and A. S. Nemirovskiĭ, *Robust Optimization*. Princeton Series in Applied Mathematics, Princeton: Princeton University Press, 2009.

[12] I. Yang, "Wasserstein Distributionally Robust Stochastic Control: A Data-Driven Approach," *arXiv:1812.09808*, Dec. 2018.

[13] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, pp. 115–166, Sept. 2018.

[14] B. Van Parys, D. Kuhn, P. Goulart, and M. Morari, "Distributionally Robust Control of Constrained Stochastic Systems," *IEEE Transactions on Automatic Control*, pp. 1–1, 2015.

[15] K. Uğurlu, "Robust optimal control using conditional risk mappings in infinite horizon," *Journal of Computational and Applied Mathematics*, vol. 344, pp. 275–287, Dec. 2018.

[16] V. Krishnamurthy, *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambridge: Cambridge University Press, 2016.

[17] M. Schuurmans, P. Sopasakis, and P. Patrinos, "Safe Learning-Based Control of Stochastic Jump Linear Systems: A Distributionally Robust Approach," in *58th IEEE Conference on Decision and Control (CDC)*, pp. 6498–6503, Dec. 2019.

[18] A. Ruszczyński, "Risk-averse dynamic programming for Markov decision processes," *Mathematical Programming*, vol. 125, pp. 235–261, Oct. 2010.

[19] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[20] A. Nemirovski, "On safe tractable approximations of chance constraints," *European Journal of Operational Research*, vol. 219, no. 3, pp. 707–718, 2012.

[21] G. C. Pflug and A. Pichler, *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering, Cham: Springer International Publishing, 2014.

[22] B. Kouvaritakis and M. Cannon, *Model Predictive Control*. Advanced Textbooks in Control and Signal Processing, Cham: Springer International Publishing, 2016.

[23] 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 *IFAC 2020 World Congress*, (Berlin), pp. 15337–15342, 2020.

[24] M. Korda, R. Gondhalekar, J. Cigler, and F. Oldewurtel, "Strongly feasible stochastic model predictive control," in *50th IEEE Conference on Decision and Control and European Control Conference*, pp. 1245–1251, Dec. 2011.

[25] M. Schuurmans and P. Patrinos, "Learning-Based Distributionally Robust Model Predictive Control of Markovian Switching Systems with Guaranteed Stability and Recursive Feasibility," *arXiv:2009.04422*, Sept. 2020.

APPENDIX

Proof of Lemma IV.4.

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 (17), for some initial state $z_0 \in \mathbf{dom}\, V$. For ease of notation, let us define $V_t := V(z_t), t \in \mathbb{N}$. Due to nonnegativity of V,

$$\mathbb{E}\left[\sum_{t=0}^{k-1} c\|x_t\|^2\right] \leq \mathbb{E}\left[V_k + \sum_{t=0}^{k-1} c\|x_t\|^2\right] = \mathbb{E}\left[V_k - V_0 + \sum_{t=0}^{k-1} c\|x_t\|^2\right] + V_0,$$

where the second equality follows from the fact that V_0 is deterministic. By linearity of the expectation, we can in turn write

$$\mathbb{E}\left[V_k - V_0 + c\sum_{t=0}^{k-1}\|x_t\|^2\right] = \mathbb{E}\left[\sum_{t=0}^{k-1} V_{t+1} - V_t + c\|x_t\|^2\right] = \sum_{t=0}^{k-1}\mathbb{E}\left[V_{t+1} - V_t + c\|x_t\|^2\right].$$

Therefore,

$$\mathbb{E}\left[c\sum_{t=0}^{k-1}\|x_t\|^2\right] - V_0 \leq \sum_{t=0}^{k-1}\mathbb{E}\left[V_{t+1} - V_t\right] + c\mathbb{E}\left[\|x_t\|^2\right]. \tag{18}$$

Recall that β^0 denotes the coordinate of β corresponding to the risk measures in the cost function (13). Defining the event $E_t := \{\omega \in \Omega \mid P_{w_t(\omega)} \in \mathcal{A}_{\beta_t^0}(s_t(\omega), w_t(\omega))\}$, and its complement $\neg E_t = \Omega \setminus E_t$, we can use the law of total expectation to write

$$\mathbb{E}\left[V_{t+1} - V_t\right] = \mathbb{E}\left[V_{t+1} - V_t \mid E_t\right]\mathbb{P}[E_t] + \mathbb{E}\left[V_{t+1} - V_t \mid \neg E_t\right]\mathbb{P}[\neg E_t].$$

By condition (5), $\mathbb{P}[\neg E_t] < \beta_t^0$. From condition (i), it follows that $z_t \in \mathbf{dom}\, V$, $\forall t \in \mathbb{N}_{[0,k]}$ and that there exists a $\bar{V} \geq 0$ such that $V(z) \leq \bar{V}$, for all $z \in \mathbf{dom}\, V$. Therefore, $\mathbb{E}[V_{t+1} - V_t \mid \neg \mathbb{E}_t] \leq \bar{V}$. Finally, by condition (ii), $\mathbb{E}\left[V_{t+1} - V_t \mid E_t\right] \leq \mathbb{E}[-c\|x_t\|^2]$. Thus,

$$\mathbb{E}\left[V_{t+1} - V_t\right] \leq \mathbb{E}\left[-c\|x_t\|^2 \mid E_t\right]\mathbb{P}[E_t] + \bar{V}\beta_t^0.$$

This allows us to simplify expression (18) as

$$\begin{aligned}
&\mathbb{E}\left[c\sum_{t=0}^{k-1}\|x_t\|^2\right] - V_0 \\
&\leq \sum_{t=0}^{k-1} -c\mathbb{E}\left[\|x_t\|^2 \mid E_t\right]\mathbb{P}[E_t] + \bar{V}\beta_t^0 + c\mathbb{E}\left[\|x_t\|^2\right] \\
&\leq \sum_{t=0}^{k-1} -c\mathbb{E}\left[\|x_t\|^2 \mid E_t\right]\mathbb{P}[E_t] + \bar{V}\beta_t^0 + \\
&\qquad c\mathbb{E}\left[\|x_t\|^2 \mid E_t\right]\mathbb{P}[E_t] + c\mathbb{E}\left[\|x_t\|^2 \mid \neg E_t\right]\mathbb{P}[\neg E_t] \\
&= \sum_{t=0}^{k-1} \bar{V}\beta_t^0 + c\mathbb{E}\left[\|x_t\|^2 \mid \neg E_t\right]\mathbb{P}[\neg E_t] \\
&\leq \sum_{t=0}^{k-1} \beta_t^0(\bar{V} + c\mathbb{E}\left[\|x_t\|^2 \mid \neg E_t\right]).
\end{aligned}$$

Since $x_t \in \mathcal{X}, t \in \mathbb{N}$, and $\mathcal{X}$ is a compact set containing the origin, there exists an $r \geq 0$ such that $\|x\|^2 \leq r$. Therefore,

$$\mathbb{E}\left[\sum_{t=0}^{k-1}\|x_t\|^2\right] \leq \tfrac{V_0}{c} + \left(\tfrac{\bar{V}}{c} + r\right)\sum_{t=0}^{k-1}\beta_t^0,$$

which remains finite as $k \to \infty$, since $(\beta_t^0)_{t\in\mathbb{N}}$ is summable. Thus, necessarily, $\lim_{t\to\infty}\mathbb{E}[\|x_t\|^2] = 0$. □

Proof of Theorem IV.5.

First, note that using the monotonicity of coherent risk measures [19, Sec. 6.3, (R2)], a straightforward inductive argument allows us to show that under Condition (i),

$$\mathbf{T}\hat{V}_N \leq \hat{V}_N, \quad \forall N \in \mathbb{N}. \tag{19}$$

Since $\bar{\mathcal{Z}} \subseteq \mathbf{dom}\, \hat{V}_N$, recall that by definition (16), we have for any $z \in \bar{\mathcal{Z}}$ that

$$\hat{V}_N(z) = \ell(x, \hat{\kappa}_N(z), w) + \rho_{w,s}^{\beta}\left[\hat{V}_{N-1}\left(\tilde{f}^{\hat{\kappa}_N}(z, v)\right)\right].$$

Therefore, we may write

$$\begin{aligned}
&\rho_{w,s}^{\beta}\left[\hat{V}_N(\tilde{f}^{\hat{\kappa}_N}(z,v))\right] - \hat{V}_N(z) \\
&= \rho_{w,s}^{\beta}\left[\hat{V}_N(\tilde{f}^{\hat{\kappa}_N}(z,v))\right] - \ell(x, \hat{\kappa}_N(z), w) \\
&- \rho_{w,s}^{\beta}\left[\hat{V}_{N-1}\left(\tilde{f}^{\hat{\kappa}_N}(z,v)\right)\right] \leq -\ell(x, \hat{\kappa}_N(z), w) \leq -c\|x\|^2,
\end{aligned}$$

where the first inequality follows by (19) and monotonicity of coherent risk measures. The second inequality follows from Condition (ii). Therefore, $V : z \to \hat{V}_N(z) + \delta_{\bar{\mathcal{Z}}}(z)$ satisfies the conditions of Lemma IV.4 and the assertion follows. □