Stable Markov Decision Processes using simulation based predictive control

Zhe Yang, Nikolas Kantas, Andrea Lecchini-Visintini, Jan M. Maciejowski

***Abstract*— In this paper we investigate the use of Model Predictive control for Markov Decision Processes under weak assumptions. We provide conditions for stability based on optimality of a specific class of cost functions. These results are useful from both a theoretical and computational perspective. When nonlinear non-Gaussian models for general state spaces are considered, the absence of analytical tools makes the use of simulation based methods necessary. Popular simulation based methods like stochastic programming and Markov Chain Monte Carlo can be used to provide open loop estimates of the optimisers. With this in mind we provide conditions under which such an approach would yield stable Markov Decision Processes.**

I. Introduction

We will consider discrete time Markov Decision Processes (MDP) defined on general state spaces; see [7] for a book length review. From a probabilistic perspective MDPs are no more than a special class of Markov Chains and can be analysed using standard results from Markov chains theory [16]. In this paper we are interested in investigating a pragmatic class of sub-optimal policies, namely Model Predictive Control (MPC) [12], [21] also known as MDPs with a rolling horizon strategy [1], [11], [23]. We aim to examine the stability properties based on appropriate performance criteria without invoking unnecessarily restrictive assumptions commonly employed, such as finite [1], [11] or countable [23] state spaces, bounded stage costs [8], linear dynamics [5], [20][5], [20] or Gaussian noise sequences [2].

Following earlier work in the field, MPC can be interpreted as an approximate Value Iteration Algorithm (VIA) [3], [4], [8], [10] where the infinite horizon cost to go is approximated by a finite horizon truncation. Much theoretical analysis comparing MPC with optimal VIA has appeared in parallel to that for deterministic non-linear MPC (for example see [18], [22]). In particular we refer the reader to [1], [11] for finite state spaces, [23] for countable ones and [8], [10] for general (measurable) state spaces. Recently the interest for stochastic MPC has been renewed. In [6] the results of [8] have been extended for problems involving the hitting and return times of some recurrent target set. In addition, in [2] MPC is treated as a particular value iteration algorithm that is implemented online, i.e. in real time or closed loop as it is commonly referred to in automatic control terminology. The author considers a specific class of problems where certainty equivalence holds.

For a detailed probabilistic analysis of the Markov chains resulting from policies computed by value and policy iteration algorithms in general state spaces we refer the reader to [15] and the references therein. In [15] the author summarises necessary stability and ergodicity assumptions required to analyse algorithms based on VIA using Lyapunov based tools such as drift conditions and the Poisson equation [17]. We will build on that framework to analyse stochastic MPC. We will consider problems where the aim is to drive the state to a specified compact target set [6]. Then using an appropriate cost formulation and assumptions similar to those found in [2], [22] one can asymptotically drive the controlled process to the target set almost surely via a sequence of nested level sets. We will see that in this sense MPC retains similar strong stability properties found in deterministic control settings [21], [12] and we will provide a discussion of how our results relate to the ones found in the literature.

It is clear that in most cases these problems cannot be computed analytically, so there is a need to employ numerical or simulation based methods for the underlying minimisations. This motivates much of the framework used as we are interested eventually to examine performance and stability when simulation based stochastic optimisation methods are employed. When simulation based methods are used for an open loop problem, MPC can generate feedback policies that are very easy to implement. In the stochastic programming literature some methods have been proposed that can find, in a known finite-number of steps, a solution to expected value criterion optimisation within the desired level of approximation and a desired confidence, while using a finite number of simulations [24], [19], [25]. These types of algorithms with probabilistic guarantees have recently been extended to include the family of Markov Chain Monte Carlo (MCMC) algorithms in [13], [14]. These guarantees provide quantitative and formal description of the behaviour of the numerical approximations resulting from different algorithms using simulation. Therefore we aim to show how this behaviour is extended when used in closed loop under appropriate MPC policies. This can provide practitioners good qualitative insight when selecting stage costs, computational resources and tools.

A. Notation

We will be using the following notation: for any scalars or vector a_i we denote $a_{1:n-1} = (a_1, \ldots, a_{n-1})$. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. We will denote the state space as

This work was supported by the European Commission under project iFly FP6- TREN-037180 and EPSRC, Grant EP/H021558/1.

Z. Yang and A. Lecchini-Visintini are with the Dept. of Engineering, University of Leicester, University Road Leicester LE1 7RH, UK {zy15,alv1}@leicester.ac.uk.

N. Kantas and J.M.Maciejowski are with the Dept. of Engineering, University of Cambridge, Trumpington St., CB21PZ, UK {nk234,jmm}@eng.cam.ac.uk.

$\mathbb{X}$, $\mathcal{X}$ the countably generated Borel sigma algebra on $\mathbb{X}$, $\mathbb{B}$ the space of Borel measureable functions and $\mathcal{P}(\mathbb{X})$ the set of probability measures on $(\mathbb{X}, \mathcal{X})$. Let also dx be the infinitesimal neighbourhood of $x \in \mathbb{X}$, $\mathbb{I}_A(x)$ the indicator function on set $A \in \mathcal{X}$ and $\delta_a(dx)$ the Dirac delta measure centered at $x = a$. Let also $L^\infty = L^\infty(\Omega, \mathscr{F}, \mathbb{P})$ be the the set of measurable functions f with finite L^∞-norm $\|f\|_\infty = \inf\{\lambda : |f| \leq \lambda\}$. A function f is called norm like if the level set $Z_\eta = \{x : f(x) \leq \eta\}$ is pre-compact for each $\eta > 0$ and $f(x) \to \infty$ as $|x| \to \infty$. For $A \in \mathcal{X}$ we will denote $P(x, A)$ to be a Markov probability transition kernel and define the semigroup's n-fold iterates $P^n(x, A) = \int P(x, dy)P^{n-1}(y, A)$ with $P^1 = P$. A kernel P (or P^n resp.) is called weakly Feller when $Pf \in L^\infty$ (or $P^n f \in L^\infty$ resp.) for $f \in L^\infty$. Also define the resolvent kernel for $\delta \in (0, 1)$ to be $K_\delta = (1-\delta)\sum_{n\geq 1} \delta^n P^n$ and the drift of the Markov chain as $\Delta V = PV - V$. For a probability measure $\mu \in \mathcal{P}(\mathbb{X})$ and a measureable function f, let $\text{supp}(\mu) = \{x : \mu(x) > 0\}$ (and similarly for f,) $\mu(f) = \int f(x)\mu(dx)$, $\mu P\{A\} = \int \mu(dx)P(x, A)$ and $P(f)(x) = \int P(x, dy)f(y)$. In addition a set Z is called petite for some probability measure $v \in \mathcal{P}(\mathbb{X})$ and some constant $\delta > 0$ if for any $x \in Z, Y \in \mathcal{X}$ we have $P(x, Y) \geq \delta\nu(Y)$.

II. Problem formulation

Let $\{X_n\}_{n\geq 0}$ be a perfectly observed $(\mathbb{X}, \mathcal{X})$ measurable Markov chain, $X_0 \sim \upsilon$ such that at time step n,

$$\begin{aligned}\mathbb{P}(X_n \in dx_n \mid X_{0:n-1} = x_{0:n-1}, A_{1:n-1} = a_{1:n-1}) \\ = P_{a_n}(x_{n-1}, dx_n),\end{aligned}$$

where $P_{a_n}(x_{n-1}, dx_n)$ is a weakly Feller probability transition kernel depending each time n on the exogenous control or action input $a_n \in \mathbb{A}$. Associated with each state $x \in \mathbb{X}$ is a compact non empty subset $\mathbb{A}(x) \subseteq \mathbb{A}$, whose elements are admissible actions or controls when the state X_{n-1} takes x as its realisation. In this and the following sections we will consider the class of admissible non-randomised Markov policy functions $\pi_n : \mathbb{X} \to \mathbb{A}$ such that $a_n = \pi_n(x_{n-1}) \in \mathbb{A}(x_{n-1})$. A sequence of functions $\pi = \{\pi_n\}_{n\geq 0}$ is usually referred as policy or feedback law. Note in this case we consider policies where the previous observed state x_{n-1} can summarise all the necessary information about the chain's previous history $a_{1:n-1}, x_{0:n-1}$. In this and the following sections we restrict the treatment to non-randomised settings. This is without loss of generality, as we can replace the image set with $\mathcal{P}(\mathbb{A}(x))$, the set of probability measures on $\mathbb{A}(x)$, and obtain a kernel for the action as for the state while still having a valid MDP. We will put this modification in use later in Section IV.

A. Open loop optimal control for non stationary problems

In order to measure the performance of an admissible policy, at each time n with $x_{n-1} = x$, we define suitable measurable non-negative stage cost functions $g_n : \mathbb{X} \times \mathbb{A} \to \mathbb{R}^+$and the expected m-stage finite horizon cost as

$$\begin{aligned}J_{n,m}(x, a_{n:n+m-1}) &:= \mathbb{E}_x\left[\sum_{k=n}^{n+m-1} g_k(X_k, a_k)\right] \\ &= \int\left[\sum_{k=n}^{n+m-1} g_k(x_k, a_k)\right]\prod_{k=n}^{n+m-1} P_{a_k}(x_{k-1}, dx_k)\delta_x(dx_{n-1}) \\ &= \sum_{k=n}^{n+m-1}\int g_k(x_k, a_k)\prod_{l=n}^{k} P_{a_l}(x_{l-1}, dx_l)\delta_x(dx_{n-1}) \\ &= \sum_{k=n}^{n+m-1} P^{k-n+1}_{a_{n:k}}(g_k)(x).\end{aligned}$$

Note that by assuming measureability with respect to g_k we are allowed to interchange the integral and summation operator[1]. Also the subscript on the $k-n+1$-order kernel P^{k-n+1} denotes its explicit dependence on the prior sequence of actions $a_{n:k}$.

An m-stage open loop policy (or equivalently control law) $\pi_{n,m}(x)$ is a policy such that at time n one computes a sequence of controls $a'_{n:n+m-1}$, which are meant to be applied successively regardless of the observations of the subsequent states. The optimal m state open-loop sequence policy at time n is denoted as $\pi^*_{n,m}(x) := a^*_{n:n+m-1}$ where

$$\pi^*_{n,m}(x) := \arg\min_{a_{n:n+m-1}\in\mathbb{A}(x)^{n-m+1}} J_{n,m}(x, a_{n:n+m-1}). \tag{1}$$

and the associated open loop optimal cost as $J^*_{n,m}(x) = J_{n,m}(x, a^*_{n:n+m-1})$.

The Bellman equation for the finite horizon problem is given by

$$J^*_{n,m}(x) = \min_{a\in\mathbb{A}(x)}\left[P_a\left(g_n + J^*_{n+1,m-1}\right)(x)\right], \tag{2}$$

the associated dynamic programming operator $T^n : \mathbb{B} \to \mathbb{B}$ on an arbitrary $P_a(x, \cdot)$-measurable function $S(x)$ is defined at time n as

$$T^n S(x) = \min_{a\in\mathbb{A}(x)}\left[P_a\left(g_n + S\right)(x)\right], \tag{3}$$

and similarly we also define for a policy function $\pi_n(x_{n-1})$ the value function operator

$$T^n_{\pi_n} S(x) = P_{\pi_n}\left(g_n + S\right)(x).$$

We also adopt the notation $T^\infty = \lim_{n\to\infty} T^1 \cdots T^{n-1}T^n S$ and similarly $T_\pi = \lim_{n\to\infty} T^1_{\pi_1} \cdots T^n_{\pi_n} S$.

In an infinite horizon setting the optimal control problem consists of minimising $J_{1,\infty}(x)$, where the state starts initially at an arbitrary $x \in \mathbb{X}$. Naturally, we have to assume that at least when the optimal open loop controls are used, the minimised cost is well defined in the sense that $\sum_{k\geq n} g_k(x_k, a^*_k)$ is $P^k_{a^*_{1:k}}$-measureable for every $k \geq 1$ including the infinite horizon case:

Assumption 1: (**A1** Regularity of minima) Suppose for all $|x| < \infty$ the underlying minimisation in $T^n J^*_{n+1,\infty}(x)$ exists for all n with the open loop minimiser being π^*_n and

$$J^*_{1,\infty}(x) = T^1 J^*_{2,\infty}(x) < \infty. \tag{4}$$

[1]This a direct application of the Lebesgue monotone convergence theorem

Clearly, this assumption is necessary in most realistic contexts.

We start by observing that (A1) implies $P^n_{\pi^*_{1,\infty}(n)}(g_n)(x) \overset{n\to\infty}{\longrightarrow} 0$. We proceed by presenting a useful lemma that reveals how in some settings a finite infinite horizon cost can imply at least asymptotic stability. We then show an example with some conditions for the transition kernel P_a and the sequence $\{g_n\}_{n\geq 1}$ that are sufficient for Assumption (A1) to hold.

Lemma 1: We have $g_n(X_n, a^*_n) \overset{n\to\infty}{\longrightarrow} 0$ $P^n_{a^*_{1:n}}$- a.s..

Proof: Proof follows from direct application of the Borel Cantelli lemma. Consider the sequence of measures $\left\{\phi_n := P^n_{a^*_{1:n}}(x, B)\right\}_{n\geq 1}$ and the sequence of sets $B_n = \{x \in \mathbb{X} : g_n(x, a^*_n) > \epsilon\}$ where $\epsilon > 0$. Using ϕ_n as the law of x_n, $\sum_{n\geq 1}\mathbb{P}_x(B_n) = \sum_{n\geq 1}\phi_n(B_n)$. From the Chebyshev's inequality we have $\phi_n(B_n) \leq \frac{1}{\epsilon}\phi_n(g_n)$ and therefore $\sum_{n\geq 1}\mathbb{P}_x(B_n) < \infty$. Using the Borel Cantelli lemma we conclude $\mathbb{P}_x(B_n \text{ i.o. }) = \mathbb{P}_x(\underset{n}{\cap}\underset{m\geq n}{\cup} B_n) = 0$ and therefore $g_n(X_n, a^*_n) \to 0$ ϕ_n- a.s.. ■

Corollary 1: Assume $\phi_n \to \phi$. Suppose there exists a petite set $\mathbb{K} \subset \mathbb{X}$ such that the system can be kept in $\mathbb{K}$ at no cost with some control sequence if it enters $\mathbb{K}$; that is $g_n = \mathbf{1}_{\mathbb{K}}c_n + \mathbf{1}_{\mathbb{K}^c}q_n$ with $\liminf_{n\to\infty} q_n = q$ and $q > \varepsilon > 0$, being $\phi-$measurable and continuous. $\operatorname{supp} g_n \cap \operatorname{supp}\phi \neq \emptyset$ and there exists $a' \in \mathbb{A}(x)$ such that $c_n(x, a') = 0$. Then under the open-loop controls $a^*_{0:\infty}$ the controlled Markov chain satisfies

$$\mathbb{P}_x(X_n \notin \mathbb{K} \text{ i.o.}) = 0 \tag{5}$$

Example 1: For this example suppose that $\{g_n\}_{n\geq 1}$ are strictly positive. One may achieve (A1) by setting $\sup_{x\in\mathbb{X}, a\in\mathbb{A}(x)}\left(\left|\frac{g_{k+1}(x,a)}{g_k(x,a)}\right|\right) = \gamma_k < 1$ with $\tilde{\gamma} = \sup_k \gamma_k < 1$ and then requiring $\sum_{k\geq 1}\tilde{\gamma}^k P^k_{\pi^*}(g_1)(x) = (1-\tilde{\gamma})^{-1} K_{\tilde{\gamma}}(g_1)(x) < \infty$ for every $|x| < \infty$. Since the integral of g_1 with the resolvent kernel is finite we recover standard regularity results for uncontrolled Markov chains [16]. In this example we place the requirement that the optimally controlled Markov chain is g_1-regular, in addition to ψ-irreducible and positive recurrent. This is a strong stability assumption and simply requiring the process to be Feller would not be adequate. In terms of stochastic stability and Lyapunov functions, a necessary and sufficient drift condition would be that for $g_1 : \mathbb{X}\times\mathbb{A} \to [1, \infty]$ there exists a petite or compact set C, constant $b < \infty$ and a non negative extended real valued function V such that for some x', $V(x') < \infty$, and so that the following holds for every $x \in \mathbb{X}$:

$$\Delta V(x) \leq -g_1(x, a') + b\mathbb{I}_C(x).$$

If g_1 is norm-like then it is enough to verify this for $P_{a'}$ and $C = \{x : g_1(x, a') \leq z\}$, where $a' \in \mathbb{A}(x)$ and $|z| < \infty$ are arbitrary.

Remark 1: The usual cost conventions for the infinite horizon are to use a discounted cost or an average cost [3], [4], [7]. Relating to the previous example in the first case we would have $\gamma_i = \gamma_j = \tilde{\gamma}$ for all i, j and in the second case we would have to use $\gamma_k = 1$ for all k. Note that the previous discussion is still relevant as regularity and Feller assumptions need to be imposed for the same reasons [15] since we would need $\frac{1}{n}P^n_\pi(g)(x) \overset{n\to\infty}{\longrightarrow} 0$ with g being the one stage cost all the time.

B. Level sets based cost formulation

The exposition so far has adopted a nonstationary cost framework. This has the crucial limitation that one cannot derive a stationary dynamic programming operator, say T, which generates the optimal value function as the fixed point of the equation $TJ^* = J^*$. This is a standard formulation in the literature and the equation appears under the name of average or discounted cost optimality equation according to which cost formulation is used [4]. Conditions under which recursive algorithms converge to such a fixed point are standard and extremely valuable [4], [9], [15]. The main reason for abandoning such a classical stationary setup is that we are primarily interested in driving the process to a desired compact target set. In contrast to [6] we will assume that this can be done via a sequence of intermediate level sets, for each of which time varying stage costs are used to penalise deviation from the level set. In this setup we no longer possess the stochastic shortest path formulation of [2] since in principle the penalty for deviating from the level sets can be made harsher as we get closer to the target set.

Our work is motivated from problems where we wish to drive the state to a compact target set $\mathbb{K}$ via a sequence of intermediate level sets $\mathbb{K}_i$. Suppose one can construct a sequence of nested sets

$$\mathbb{K} \subset \ldots \subset \mathbb{K}_n \subset \ldots \subset \mathbb{K}_2 \subset \mathbb{K}_1 \subset \mathbb{X}, \tag{6}$$
$$\text{where } \mathbb{K}_i = \{x : V(x) \leq z_i\} \text{and } z_1 > z_2 > \ldots > z_n$$

with V being an arbitrary norm-like *potential fuction*. In the spirit of Corollary1 we will focus our attention on cost functions of the form $g_n = \mathbf{1}_{\mathbb{K}_n}c_n + \mathbf{1}_{\mathbb{K}^c_n}q_n$ with each q_n, c_n norm-like, bounded below by $\gamma_n^{-1}V$ and above by $\gamma_n V$ for $\gamma_n \in (1, \infty)$, and such that $q_n(x') = c_n(x')$ when x' lies on the boundary of $\mathbb{K}_n$ to ensure continuity of g_n. As in Corollary 1 we minimise the infinite horizon problem with the objective being to drive the state eventually to the target set $\mathbb{K}$ via a sequence of intermediate level sets. We formalise this setup with the following assumption:

Assumption 2: (**A2** Level sets construction) Assume that $V(x)$ is a normlike potential function and set $g_n = \mathbf{1}_{\mathbb{K}_n}c_n + \mathbf{1}_{\mathbb{K}^c_n}q_n$, with $\mathbb{K}_i$ as in (6), $q_n(x', a) = c_n(x', a)$ when $x' \in \partial\mathbb{K}_n$, $\liminf_{n\to\infty} c_n = 0$, $\liminf_{n\to\infty} q_n = \varepsilon > 0$, and for each n there exist $\gamma_n \in (1, \infty)$ such that $\gamma_n^{-1}V \leq g_n \leq \gamma_n V$ with $\limsup_{n\to\infty}\gamma_n \geq 1$.

Note that a different cost is used inside the level set than that outside. An analogous dual mode construction has appeared in a deterministic control setting in [18] where the authors assume that the process follows a particular policy inside the target set and a MPC policy is used to force the process to return when the process escapes the target set.

In our formulation the policy is defined by minimising a different cost inside and outside the level sets.

C. Ideal closed loop optimal control and Model Predictive Control

Applying an open loop solution of control sequences has the obvious disadvantage that it does not take into account that future observations of X_n will become available. Nevertheless an open loop solution can be still used as a base policy in conjunction with dynamic programming [3]. We will refer to the closed loop policy or feedback law being a policy that uses the realised observations of the process. Clearly any Markov policy as described earlier consists of a closed loop policy.

Assuming it is possible and that the minimisers exist, ideally one may wish to implement a feedback strategy to solve $T^n J^*_{n+1,\infty}$ for every n. We will denote this ideal Markov policy as $\pi^{CL} = \{\pi^*_{n,\infty}(x_{n-1})(1)\}_{n\geq 1}$ where $\pi^*_{n,\infty}(x)(1) = \arg\min_{a\in\mathbb{A}(x)}\left[P_a\left(g_n + J^*_{n+1,\infty}\right)(x)\right]$. Even in the ideal case where this is possible, an important requirement that (2) can be solved for every n, which is indeed assumed in (A1). Clearly this ideal feedback policy is in most cases very hard if not impossible to implement. Instead we shall resort to MPC and treat the corresponding value function as a truncated approximation of the ideal one. Assuming at time n we have observed the previous state to be x_{n-1}, an MPC or receding horizon policy uses only the first element of the computed open loop control sequence, i.e. $a^\star_n = \pi^*_{n,h}(x_{n-1})(1)$. After $a^\star_n$ is applied and x_n is observed the procedure is repeated. For a fixed lookahead horizon an MPC policy can be written as $\pi^{MPC} = \{\pi^*_{n,h}(x_{n-1})(1)\}_{n\geq 1}$. Note that this will be a stationary policy if the prediction horizon h remains the same at each n.

We will impose that the additional assumption holds:

Assumption 3: (**A3** Foster-Lyapunov condition) Assume the following drift condition holds for the potential V such that for each n and some $a = \pi^*_{n,h}(x)(1)$ with (possibly large) $h > 1$,

$$P_a V(x) \leq \lambda V(x) \tag{7}$$

where $\lambda \in (0,1)$

This is in fact a uniform ergodicity assumption [16], which also implies the existence of a unique invariant measure ϕ. The assumption is quite strong, but seems to be common when working with non-stationary or unbounded costs [4], [9], [10], [15]. Note that since P_a is weakly Feller and V norm-like equation (7) and Assumption A3 can be in practice verified only in compact state spaces, where the existence of a positive probability density function is imposed, otherwise one should can relax (A3) and require the same condition for an appropriate resolvent K_δ instead (see [8], [10], [15].)

To determine stochastic stability of iterative schemes based on VIA ideas, such as MPC, we intend to show that these iterative algorithms are stable in the sense that they yield a finite cost at the infinite time horizon. Then by a straightforward application of (7) the resulting Markov chain will be V- regular or geometrically regular [16], [17]. In this sense we are no longer interested in investigating how the resulting suboptimal cost compares to the optimal infinite horizon cost as done in [6], [8], [10]. This motivates further discarding stationary value iteration schemes as the objective is not to assert whether a recursive algorithm based on VIA for stationary problems will converge to a (possibly unique) optimal solution as in [4], [6], [8], [10], [15]. Instead, the particular choice of cost ensures that when the total accumulated infinite horizon cost of the Markov chain is finite, then the process will asymptotically reach the target set almost surely.

III. Model predictive control and stochastic stability

We summarise the algorithm to generate the MPC policy $\pi^{MPC} = \{\pi^*_{n,h}(x_{n-1})(1)\}_{n\geq 1} = \{a^\star_n\}_{n\geq 1}$as follows:

- Initialise x_0. For $n \geq 1$
  - Compute $\pi^*_{n,h}(x_{n-1})$, $J^*_{n,h}(x_{n-1})$ such that $J_{n,h}(x_{n-1}, \pi^*_{n,h}(x_{n-1})) = T^n \cdots T^{n+h-1}\mathbf{0}$.
  - Return $a^\star_n = \pi^*_{n,h}(x_{n-1})(1)$ and sample $x_n \sim P_{a^\star_n}(x_{n-1},\cdot)$

At each time n,

$$V_n(x) = \sum_{k=n+1}^{n+h-1} P^{k-n+1}_{\pi^*_{n,h}(1),\ldots,\pi^*_{n,h}(k)}(g_k)(x_{n-1})$$

will act as the value function approximation[2] of $P_{\pi^*_{n,\infty}(1)} J^*_{n+1,\infty}(x_n)$ resulting from the truncation due to the prediction horizon. The feedback policy is generated by repeatedly computing an optimal finite horizon problem in a receding horizon manner. As with the open loop case, we will denote the expected j-stage closed-loop cost as

$$J^\star_{n,m,j}(x) := \mathbb{E}_x\left[\sum_{k=n}^{n+j-1} g_k(X_k, a^\star_k)\right], \tag{8}$$

where the process starts from state x and the closed loop action sequence is given by $\{a^\star_n\}_{n\geq 1}$.

In the remainder of this section we will list some basic results concerning the asymptotic stability of stochastic MPC for the given model. The results are novel and can be viewed as an extension of the deterministic counterparts found in [2], [22] for Markov chains defined on general state spaces. Compared to related results of [6] we use of unbounded costs and present the analysis of MPC as a forward only approximate dynamic programming implementation. Also we do not restrict g_n to be in some sense strictly decreasing. Instead we allow some oscillation and assume that the particular MPC prediction horizon naturally selects a non increasing sub-sequence for $P^h_{a^*_{n:n+h}}(g_{n+h})$. In detail, we pose the following assumption:

Assumption 4: Assume at least one of the following is true:

[2]In the subscript of the $k-n$-fold kernel we use the convention that for $k=1$ then $\pi^*_{n,h}(1),\ldots,\pi^*_{n,h}(k)$ is simply $\pi^*_{n,h}(1)$.

i) There exist h such that for every n and x with $|V(x)| < \infty$ the following holds

$$J^*_{n,h}(x) \leq J^*_{n,h-1}(x) + b_n, \tag{9}$$

where $\sum_{n\geq 1} b_n < \infty$.

ii) There exist h and $\delta \in (0,1)$ such that for every n and x with $|V(x)| < \infty$ we have:

$$\min_{a\in A(x)} P^h_{\pi^*_{n,h}} P_a\left(g_{n+h}\right)(x) \leq \delta P_{\pi^*_{n,h}(1)}\left(g_n\right)(x)$$

where the minimisation is supposed to exist when $\pi^*_{n,h}$ is the sequence of actions to minimise $J_{n.h}$ as in (1).

Proposition 1: If Assumptions (A1-4) hold, then for every n and x with $|V(x)| < \infty$ we have:

$$J^\star_{n,h,\infty}(x) < \infty. \tag{10}$$

Proof: We will split the proof in two parts showing (A1-3) and Assumption 4 i) (and ii) respectively) $\Rightarrow$ Proposition 1.

Part i): At time n consider the Bellman's optimality condition (2) for

$$J^*_{n,h}(x') = P_{a^*} g_n(x') + P_{a^*} J^*_{n+1,h-1}(x'),$$

where $a^* = \pi^*_{n,h}(1)$. Using (9) we get

$$J^*_{n,h}(x') \geq P_{a^*} g_n(x') + P_{a^*} J^*_{n+1,h}(x') - b_{n+1},$$

and by integrating both sides with the law of the path of the chain up to time n, $L_{n-1}(x,x') = P^{n-1}_{a^\star_{1:n-1}}(x,x')$ we get

$$L_{n-1}P_{a^*}\left(g_n\right)(x) \leq L_{n-1}\left(J^*_{n,h} - P_{a^*}\left(J^*_{n+1,h}\right)\right)(x) + b_{n+1}.$$

We may omit the initial condition x without confusion and consider the telescopic sum for every n:

$$\sum_{n\geq 1} L_{n-1}P_{a^*}\left(g_n\right) \leq \sum_{n\geq 1} L_{n-1}\left(J^*_{n,h} - P_{a^*}\left(J^*_{n+1,h}\right)\right) + \sum_{n\geq 1} b_n. \tag{11}$$

For the lhs term we identify $\sum_{n\geq 1} L_{n-1}P_{a^*}\left(g_n\right) = J^\star_{n,h,\infty}(x)$ and for first term of the rhs we use (A3) to construct the following bounds

$$J^*_{n,h}(x') \leq V(x') \sum_{k=n}^{n+h-1} \gamma_k \lambda^{k-n+1} \tag{12}$$

and hence

$$\sum_{n\geq 1} L_{n-1} J^*_{n,h}(x) \leq V(x) C_n$$

where $C_n = \sum_{n\geq 1}\lambda^n \sum_{k=n}^{n+h-1}\gamma_k\lambda^{k-n+1} < \infty$. Given the each term in the telescopic sum is converging in the set defined by $V(x) < \infty$ we can use appropriate cancellations in (11) to get

$$\sum_{n\geq 1} L_{n-1}\left(J^*_{n,h} - P_{a^*}\left(J^*_{n+1,h}\right)\right) \leq J^*_{1,\infty} < \infty. \tag{13}$$

For the second term in the rhs (11) we clearly have $\sum_{n\geq 1} b_{n+1} < \infty$ and therefore reach the desired result.

Part ii): This time at time $n+1$ we construct a policy comprised of the last $h-1$ elements of $\pi^*_{n,h}$ followed by a specific action $\check{\alpha}$, i.e. we let $\tilde{\pi}_{n+1} = \left(\pi^*_{n,h}(2),\ldots,\pi^*_{n,h}(h),\check{\alpha}\right)$ and $a^* = \pi^*_{n,h}(1)$. We specifically choose $\check{\alpha} = \arg\min_{a\in A(x)} P^h_{\pi^*_{n,h}} P_a\left(g_{n+h}\right)(x)$. We will compare the resulting finite horizon cost $J_{n+1,h}(y,\tilde{\pi}_{n+1})$ at time $n+1$, where $J_{n+1,h}(y,\tilde{\pi}_{n+1}) =$

$$\sum_{k=n+1}^{n+h} P^{k-n+1}_{\tilde{\pi}_{n+1}(1),\ldots,\tilde{\pi}_{n+1}(k)}\left(g_k\right)(y) = T^{n+1}_{\tilde{\pi}_{n+1}(1)}\cdots T^{n+h}_{\tilde{\pi}_{n+1}(h)}(y),$$

with the optimal cost $J^*_{n+1,h}(y)$. By the principle of optimality and after integrating appropriately withP_{a^*}, we have $P_{a^*}T^{n+1}_{\tilde{\pi}_{n+1}(1)}\cdots T^{n+h}_{\tilde{\pi}_{n+1}(h)}(x') \geq P_{a^*}J^*_{n+1,h}(x')$. Hence the following decomposition holds

$$\sum_{k=n}^{n+h-1} P^{k-n+1}_{\pi^*_{n,h}(1),\ldots,\pi^*_{n,h}(h)} g_k(x') - P_{a^*}g_n(x') + P_{a^*}P^h_{\tilde{\pi}_{n+1}} g_{n+h}(x') \geq P_{a^*}J^*_{n+1,h}(x) \tag{14}$$

and if we substitute

$$J^*_{n,h}(x) = \sum_{k=n}^{n+h-1} P^{k-n+1}_{\pi^*_{n,h}(1),\ldots,\pi^*_{n,h}(h)}\left(g_k\right)(x')$$

in (14) and then also integrate both sides with the law of the path of the chain up to time n, $L_{n-1}(x,x') = P^{n-1}_{a^\star_{1:n-1}}(x,x')$ we get

$$\begin{aligned} L_{n-1}P_{a^*}\left(g_n\right)(x) &\leq L_{n-1}\left(J^*_{n,h} - P_{a^*}\left(J^*_{n+1,h}\right)\right)(x) \\ &+ L_{n-1}\left(P_{a^*}P_{\tilde{\pi}_{n+1}} g_{n+h}\right)(x). \end{aligned}$$

We may omit the initial condition x without confusion and consider the telescopic sum for every n:

$$\sum_{n\geq 1} L_{n-1}P_{a^*}\left(g_n\right) \leq \sum_{n\geq 1} L_{n-1}\left(J^*_{n,h} - P_{a^*}\left(J^*_{n+1,h}\right)\right) + \sum_{n\geq 1} L_{n-1}P_{a^*}P_{\tilde{\pi}_{n+1}}\left(g_{n+h}\right). \tag{15}$$

For the lhs term we again identify $\sum_{n\geq 1} L_{n-1}P_{a^*}\left(g_n\right) = J^\star_{n,h,\infty}(x)$ and for first term of the rhs we can use (13) again to show it is finite. For the second sum of the rhs of (15) we can show that the sum is converging within the set defined by $V(x) < \infty$ by direct consequence of drift condition (A2) and Assumption 4 ii) as follows: $\sum_{n\geq 1} L_{n-1}P_{a^*}P_{\tilde{\pi}_{n+1}}\left(g_{n+h}\right) \leq \sum_{n\geq 1} L_{n-1}\delta P_{a*}\left(g_n\right) \leq V(x)\delta\sum_{n\geq 1}\lambda^n\gamma_n$. ∎

Remark 2: Note that it is trivial to show that in case were $h \to \infty$ then equality holds and

$$\lim_{h\to\infty}\left|J^\star_{1,h,\infty}(x) - J^*_{1,\infty}(x)\right| \to 0 \tag{16}$$

Corollary 2: (cont. from Corollary 1) Let ϕ be the invariant measure of the chain. Also, for every x_0 such that $|V(x_0)| < \infty$, under the policy π^{MPC} we have $\mathbb{P}_x(X_n \notin \mathbb{K} \text{ i.o.}) = 0$.

IV. Randomised algorithms for rolling horizon policies

In this section we consider the case when a randomised open loop policy is employed. We propose to use samples obtained from stochastic optimisation algorithms as estimates of the open loop optimisers $\pi^*_{n,h}(x_{n-1})$, where throughout the section we will assume that Assumption 4 holds. In this context, existing stochastic approximation algorithms are able to provide ϵ-optimal solutions [4] with a desired statistical confidence ρ [25], utilising a number of Monte Carlo simulations which grows polynomially with ϵ^{-1} and the desired ρ [19], [13], [14]. We will view these algorithms as complex Markov transition kernels that can satisfy conditions based on two alternative ϵ-optimality notions. In the last part of this section we will comment on how this setting applies for particular simulation based algorithms.

A. ϵ-optimality

At time n consider the augmented Markov kernel $\Pi^h_n : \mathbb{X} \to \mathcal{P}\left(\mathbb{A}^h\right)$ which aims to generate samples that approximate the minimiser of $J_{n,h}(x,\alpha)$ with α or α_n denoting $a_{n:n+h-1}$ generically in this section. Also define the ϵ-optimality region at time n as

$$\mathcal{B}^h_n(\epsilon) = \{\alpha \in \mathbb{A}(x)^h : J_{n,h}(x,\alpha) \leq J^*_{n,h}(x) + \epsilon\},$$

We will assume that if $\alpha \in \mathcal{B}^h_n(\epsilon)$ then a drift condition similar to (A3) holds for its first element $\alpha(1)$:

Assumption 5: **(A3)'** Assume the following drift condition holds for the potential V such that for each n with $\alpha \in \mathcal{B}^h_n(\epsilon)$,

$$P_{\alpha(1)}V(x) \leq \lambda V(x) \tag{17}$$

where $\lambda \in (0,1)$ and h is as in Assumption 4

We want generate samples $\tilde{\pi}_{n,h} \sim \Pi^h_n(x,\cdot)$ using simulation based methods that satisfy some performance criteria based on an imprecision ϵ_n with some desired statistical confidence $\rho_n > 0$. In this sense we should be able to design a sequence ϵ_n, ρ_n such that:

$$\Pi^h_n(x, \mathcal{B}^h_n(\epsilon_n)) = \rho_n.$$

Assumption 6: For every $n \geq 1$ and x such that $V(x) < \infty$, let $\text{supp}\Pi^h_n(x,d\alpha) = \mathbb{A}(x)^h$ and assume that $\sum_{n\geq 1}\epsilon_n < \infty$ and $\prod_{n\geq 1}\rho_n > 0$.

For example the condition on ϵ_n, ρ_n can be fulfilled if ϵ_n tends polynomially fast to zero and ρ_n exponentially fast to one.

It seems we will consider the MPC policy to be an arbitrary Markov transition kernel $\Pi_n : \mathbb{X} \to \mathcal{P}(\mathbb{A})$, which is generated as the marginal of the augmented policy Π^h_n with respect to all actions apart the first one, $\Pi_n = \Pi^h_n\left(\prod_{k=n+1}^{n+h-1}\mathbb{I}_{a_k\in\mathbb{A}(x)}\right)$. Then define the transition kernel MDP's state when randomised MPC is employed as $P_n(x,dx') = \int_{\mathbb{A}(x)}\Pi_n(x,da)P_a(x,dx')$ and rewrite the j-step MPC cost to go function at time n as

$$\widetilde{J}_{n,h,j}(x) = \sum_{k=n}^{n+j-1} P^{k-n+1}_k\left(g_k\right)(x). \tag{18}$$

We will assume throughout that all the integrals with respect to Π_n, Π^h_n are well defined. As in the previous section we will show the following proposition:

Proposition 2: If Assumptions (A1-2), (A3)' , 4 and 6 hold, then for every n and x with $|V(x)| < \infty$ we have:

$$\widetilde{J}_{n,h,\infty}(x) < \infty. \tag{19}$$

Proof: Sketch of proof follows. Let $\kappa_n(x) = J^*_{n,h}(x) + \epsilon_n > 0$. We have defined $\mathcal{B}^h_n(\epsilon_n)$ as the region where $J_{n,h}(x,\alpha) \leq \mathbb{I}_{J_{n,h}(x,\alpha)\leq\kappa_n(x)}\kappa_n(x)$ holds, so by integration we get

$$\Pi^h_n(J_{n,h})(x) \leq \rho_n\kappa_n(x). \tag{20}$$

Similarly if we use (A3)' in the same way it was used to construct (12) we can show $P_{\alpha(1)}J_{n,h}(x) \leq C^1_n J_{n,h}(x)$ with $C^1_n < 1$ being a constant decreasing with n and hence we get

$$\Pi^h_n P_{\alpha(1)}(J_{n,h})(x) \leq C^1_n\rho_n\kappa_n(x). \tag{21}$$

The remaining proof follows similar arguments to those in the proof of Proposition 1. We split the proof in two parts each using Assumption 4 i) and ii).

Part i): Consider the following decomposition:

$$\begin{aligned}\Pi^h_n(J_{n,h})(x') &= P_n(g_n)(x') + \Pi^h_n P_{\alpha_n(1)}\left(J_{n+1,h-1}\right)(x') \\ &\geq P_n(g_n)(x') + \Pi^h_n P_{\alpha_n(1)}\left(J_{n+1,h}\right)(x') - b_{n+1}.\end{aligned} \tag{22}$$

By the same arguments as in Part i) of the proof of Proposition 1 we can derive the telescopic sum where x is omitted:

$$\begin{aligned}\widetilde{J}_{n,h,j}(x) \leq &\sum_{n\geq 1} L_{n-1}\left(\Pi^h_n(J_{n,h}) - \Pi^h_n P_{\alpha_n(1)}\left(J_{n+1,h}\right)\right) \\ &+ \sum_{n\geq 1} b_{n+1},\end{aligned}$$

where L_n is the appropriately modified law of the path of the Markov chain. For the first term of the rhs we observe that it is less than or equal to $\sum_{n\geq 1} C^2_n(1-C^1_n)\rho_n\kappa_n(x) < \infty$ when $V(x) < \infty$. For the sake of brevity we omit intermediate steps.

Part ii) Using similar arguments as Part i) and Part ii) of the proof of Proposition 1 we go directly to the inequality with the telescopic sum

$$\begin{aligned}\widetilde{J}_{n,h,j}(x) \leq &\sum_{n\geq 1} L_{n-1}\Pi^h_n\left(J^*_{n,h} - P_{\alpha_n(1)}\left(J^*_{n+1,h}\right)\right). \\ &+ \sum_{n\geq 1} L_{n-1}P_{\alpha_n(1)}\Pi^h_{n+1}\left(g_{n+h}\right).\end{aligned}$$

The first term in the rhs can be treated in a similar fashion as before. For the second term we may use repeatedly (A2)-(A3)' as in (20)-(21) to get the required result. ∎

A more complete proof will be presented in future versions of this paper.

Corollary 3: (cont. from Corollary 2) By adapting Lemma 1 appropriately it is possible to show that if the Markov chain X_n starts at x_0 with $|V(x_0)| < \infty$ and evolves with the transition kernel P_n, where Π_n satisfies either Assumptions 6, then we have $\mathbb{P}_x(X_n \notin \mathbb{K} \text{ i.o.}) = 0$.

B. Approximate domain optimality

In the second part of the section we will consider the following alternative probabilistic notion of ϵ-optimality based on ideas from [25] and referred to as approximate domain optimality in [13], [14]. First consider the region around α defined as

$$\mathcal{B}_n^h(\alpha,\epsilon)=\{\alpha'\in\mathbb{A}(x)^h : J_{n,h}(x,\alpha')<J_{n,h}(x,\alpha)-\epsilon\}.$$

Then for some arbitrary finite measure μ consider the so called approximate domain optimality region:

$$\mathcal{R}_n^h(\varepsilon,\chi)=\{\alpha\in\mathbb{A}(x)^h :\ \mu(\mathcal{B}_n^h(\alpha,\epsilon))\leq\chi\mu(\mathbb{A}(x)^h)\}.$$

In the same spirit as before we can specify ϵ_n, ϱ_n and χ_n to design a kernel Π_n^h satisfying

$$\Pi_n^h(x,\mathcal{R}_n^h(\epsilon_n,\chi_n))=\varrho_n$$

where $\varrho_n>0$. We pose the following assumption:

Assumption 7: For every n and x such that $V(x)<\infty$ let $\text{supp}\Pi_n^h(x,d\alpha)=\mathbb{A}(x)^h$, assume that the Radon-Nicodym derivative $\frac{d\Pi_n^h}{d\mu}$ exists and for $\epsilon_n,\varrho_n,\chi_n$ we have that $\sum_{k=0}^{\infty}\chi_k<\infty$, $\sum_{n\geq1}\epsilon_n<\infty$ and $\prod_{n\geq1}\varrho_n>0$.

Conjecture 1: If Assumptions (A1-2), (A3)', 4 and 7 hold, then for every n and x with $|V(x)|<\infty$ we have:

$$\widetilde{J}_{n,h,\infty}(x)<\infty. \tag{23}$$

We will present the proof in future versions of the paper.

C. Existing algorithms satisfying Assumptions 6, 7

Efficient algorithms equipped with provable bounds, of the type required by Assumptions 6, include stochastic programming approaches [19] and [24], which were developed for convex stochastic problems. The same type of guarantees have been shown to hold for non-convex stochastic problems in [13], [14] when Markov Chain Monte Carlo (MCMC) is employed. In all these cases, given desired values of accuracy ϵ and confidence ρ, an explicit bound on the required number of Monte Carlo simulations can be obtained. Such a bound is effectively a stopping criterion for the optimisation algorithm. However, in order to compute such a bound one requires knowledge of a bound for the Lipschitz constant of the cost. If this is not available then similar guarantees can be obtained for the weaker notion of approximate domain optimality in Assumption 7 (for more details see [13], [14]). In the latter case Lipschitz continuity is not required. It is conjectured that the stability results of this section for ϵ-optimality can be extended to the case of approximate domain optimality. This is the object of current investigation.

V. Conclusions

We have developed an approach in which we do not adopt *a-priori* a specific class of MDP. The cost considered is required to obey a stochastic Lyapunov drift condition with respect to the process transition kernel, but is otherwise left flexible with respect to the controller's design needs. Under fairly weak assumptions the optimal MPC policy was shown to exhibit some fairly general stability properties, which were established using dynamic programming and the Markovian nature of the system. In the last part of the paper we considered using simulation based methods, which can be employed for a wide class of models to obtain estimates of the MPC optimisers. In our future work we plan show examples of problems where our assumptions can be verified together with some numerical comparisons.

Acknowledgements: The authors would like to thank Ellie Siva and Tom Dean for valuable comments to improve the presentation of the paper.

References

[1] Alden J. M. and Smith R. L., Rolling Horizon Procedures in Nonhomogeneous Markov Decision Processes, Oper. Res., Vol. 40, Suppl. 2: Stochastic Processes, pp. S183-S194, 1992 .

[2] Bertsekas D. P., Dynamic Programming and Suboptimal Control: A Survey from ADP to MPC, Eur. J. of Contr., Vol. 11, Nos. 4-5, 2005

[3] Bertsekas D. P., Dynamic Programming and Optimal Control, Volumes 1 and 2, Athena Scientific, 2001.

[4] Bertsekas D.P. and Shreve S.E., Stochastic Optimal Control: The Discrete-Time Case, Athena Scientific, 1996.

[5] Cannon M., Kouvaritakis B. and Wu X., Model predictive control for systems with stochastic multiplicative uncertainty and probabilistic constraints, Automatica, 45, pp 167-172, 2009.

[6] Chatterjee D., Cinquemani E., Chaloulos G. and Lygeros J. Stochastic control up to a hitting time: optimality and rolling-horizon implementation, arXiv:0806.3008v3 [math.OC], 2009.

[7] Feinberg E. A. and Shwartz A. Handbook of Markov Decision Processes: Methods and Applications, Kluwer Int. Series, 2002.

[8] Hernandez-Lerma O. and J.B. Lasserre, Error bounds for rolling horizon policies in general Markov control processes, IEEE Trans. Auto. Contr. 35, pp 1118–1124, 1990.

[9] Hernandez-Lerma O. and J.B. Lasserre, Average cost optimal policies for Markov control processes with Borel state space and unbounded costs, Syst. Contr. Lett. 15, pp. 349–356, 1990.

[10] Hernandez-Lerma O. and J.B. Lasserre, Value Iteration and Rolling Plans for Markov control processes with unbounded rewards, J. Math. Anal. Appl. 177, pp. 38–55, 1993.

[11] Hopp W. J., Bean J. C., Smith R. L., A New Optimality Criterion for Nonhomogeneous Markov Decision Processes, Oper. Res., Vol. 35, No. 6, pp. 875-883, 1987.

[12] Maciejowski J.M., Predictive control with Constraints, Prentice Hall, 2002.

[13] Lecchini-Visintini A., Lygeros J. and Maciejowski J. M., Simulated Annealing: Rigorous finite-time guarantees for optimization on continuous domains, In Advances NIPS 27, 2008.

[14] Lecchini-Visintini A., Lygeros J. and Maciejowski J. M., Stochastic optimization on continuous domains with finite-time guarantees by Markov chain Monte Carlo methods, IEEE Trans. Aut. Contr., to appear.

[15] Meyn S., Stability, Performance evaluation and Optimisation, in [7].

[16] Meyn S. and Tweedie R.L., Markov Chains and Stochastic Stability, Springer Verlag, 1993.

[17] Meyn S. and Tweedie R.L., Stability of Markovian processes I: criteria for discrete Markov chains, Adv. Appl. Prob, 24, 542-574, 1992.

[18] Michalska H. and Mayne D. Q., Robust Receding Horizon Control of Constrained Nonlinear Systems, IEEE Trans. Aut. Contr., 38(11):1623–1632, 1993.

[19] Nesterov Y. and Vial J. P., Confidence level solutions for stochastic programming,Automatica, 44, Vol 6, pp 1559-1568, 2008.

[20] Primbs J.A. and Sung C.H., Stochastic Receding Horizon Control of Constrained Linear Systems with state and control multiplicative noise, IEEE Trans. Aut. Contr, 54(2), pp 221 - 230, 2009.

[21] Rawlings J.B. and Mayne D. Q., Model Predictive Control: Theory and Design, Madison, Wisconsin: Nob Hill Publishing, 2009.

[22] Scokaert P. O. M., Mayne D. Q. and Rawlings J. B., Suboptimal model predictive control (feasibility implies stability), IEEE Trans. Aut. Contr, 44(3):648–654, 1999.

[23] Sethi S. and Sorger G., A theory of rolling horizon decision making, Ann. Oper. Res. 29, 387-416, 1991.

[24] Shapiro, A., Stochastic programming approach to optimization under uncertainty, Math. Program., Ser. B, 112,183-220, 2008.

[25] Vidyasagar M., Randomized algorithms for robust controller synthesis using statistical learning theory, Automatica, Vol 37(10), pp 1515-1528, 2001.