Economic MPC of Markov Decision Processes: Dissipativity in Undiscounted Infinite-Horizon Optimal Control ⋆

Sebastien Gros [a], Mario Zanon [b]

[a] *Dept. of Eng. Cybernetics, Faculty of Information Technology, NTNU, Trondheim, Norway*

[b] *IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100, Lucca, Italy*

Abstract

Economic Model Predictive Control (MPC) dissipativity theory is central to discussing the stability of policies resulting from minimizing economic stage costs. In its current form, the dissipativity theory for economic MPC applies to problems based on deterministic dynamics or to very specific classes of stochastic problems, and does not readily extend to generic Markov Decision Processes. In this paper, we clarify the core reason for this difficulty, and propose a generalization of the economic MPC dissipativity theory that circumvents it. This generalization focuses on undiscounted infinite-horizon problems and is based on nonlinear stage cost functionals, allowing one to discuss the Lyapunov asymptotic stability of policies for Markov Decision Processes in terms of the probability measures underlying their stochastic dynamics. This theory is illustrated for the stochastic Linear Quadratic Regulator with Gaussian process noise, for which a storage functional can be provided explicitly. For the sake of brevity, we limit our discussion to undiscounted Markov Decision Processes.

Key words: Markov Decision Processes, dissipativity for economic MPC, storage functions, economic costs

1 Introduction

The use of optimization-based policies is widespread in control, with the notable example of Model Predictive Control (MPC) which has gained increasing popularity in the last decades. In the deterministic setting, MPC schemes are often used to steer a system to a given feasible reference input and state. In that context, the stage cost minimized in the MPC scheme is typically convex, taking its minimum at the reference. Quadratic costs are the most common choice. This type of MPC scheme is commonly referred to as *tracking MPC*.

In a control context, the system stability in closed loop with a policy is a crucial feature. In particular, an asymptotically stable policy ensures that the closed-loop system will be steered to a specific steady-state. The stability of tracking MPC schemes for deterministic systems is fairly straightforward to establish, and—under a mild controllability assumption—simply requires the MPC stage cost to be lower-bounded by a class-$\mathcal{K}_\infty$ function, with the possible addition of a terminal cost and constraint set in the finite-horizon setting. When these criteria are fulfilled, a stability argument can be easily constructed via the Lyapunov stability theory [9,19].

Tracking MPC schemes are, however, fairly restrictive as to which stage cost can be used, and using stage costs belonging to a broader class of functions can be beneficial. Indeed, the recent literature on MPC argues for the use of *economic* stage costs, representing directly the performance of the system with regard to the overall control goals, rather than the specific objective of steering the system to a given reference, see, e.g., [7,10,21]. Such economic objectives often correspond to the energy, the time or the financial cost of performing a given task. It is commonly argued that a policy minimizing an economic stage cost is more conducive to maximizing the system performance than a policy optimizing a tracking stage cost can be [18].

While appealing, economic stage costs typically do not

⋆ This paper was not presented at any IFAC meeting. Corresponding author M. Zanon. This paper was partially supported by the Italian Ministry of University and Research under the PRIN'17 project "Data-driven learning of constrained control systems" , contract no. 2017J89ARP; by ARTES 4.0 Advanced Robotics and enabling digital Technologies & Systems 4.0, CUP: B81J18000090008; and by the Norwegian Research Council project "Safe Reinforcement-Learning using MPC" (SARLEM).

Email address: `sebastien.gros@ntnu.no` (Sebastien Gros).

satisfy the criteria required to conclude the stability of the resulting optimal control policy. To address that issue, a new stability theory has been developed, commonly referred to as *dissipativity* theory in the context of economic MPC. We ought to stress here that this dissipativity framework is very specific, as it focuses on the analysis of systems in closed-loop with optimal control policies. This framework ought not be mistaken for the general dissipativity theory for dynamical systems. The key idea behind economic MPC dissipativity theory is to transform the economic stage cost into one that is lower-bounded by a class-$\mathcal{K}_\infty$ function, while leaving the resulting policy unchanged. This transformation is often referred to as *cost rotation*, and performed via a so-called *storage function*. If this transformation is possible, i.e., if strict dissipativity holds, then the stability of economic MPC can be analyzed via the Lyapunov stability theory [1,5,6,8,15,24]. A strong point of the economic MPC disspativity theory is that it is based on the interplay between stability and optimality, and therefore provides a natural way of discussing the stability of optimal policies without requiring the construction of the actual policy. The aforementioned dissipativity theory applies to systems having deterministic dynamics, and is not yet extended to general stochastic systems, with the notable exceptions of [3,20] which, however, only apply to rather specific settings.

MPC for stochastic systems is often treated within the *Robust MPC* (RMPC) or *Stochastic MPC* (SMPC) frameworks. The former is equipped with stability theories, but the analysis is usually restricted to tracking MPC formulations, to the exception of [2,3], which discuss the economic setting. However, the stability results found in the RMPC context are limited to proving the stability of the closed-loop trajectories to a set [4,12,23] under the assumption of bounded support. That set can be arbitrarily large and the behavior of the trajectories within the set is not discussed by the theory. SMPC often targets the minimization of the given stage cost in terms of its expected value taken over the stochastic predicted trajectories and the associated stability theories are less mature [13].

Minimizing the expected value of a stage cost subject to stochastic trajectories is generally referred to as a Markov Decision Process (MDP). MDPs can be formulated in a variety of settings. For the sake of brevity, we will focus on discrete-time bias-optimal undiscounted MDPs over an infinite horizon and continuous state spaces, which are the closest ones to economic MPC and to the current deterministic economic MPC dissipativity setup. The extension of our results to other settings is the subject of current research.

The stability of MDPs can be arguably analyzed in the broader context of Markov Chains [14]. Unfortunately, this framework provides results that are not easily related to optimality and therefore to the original MDPs. A discussion on the stability of MDPs in the context of economic MPC would therefore be beneficial, as they would bridge that gap. Unfortunately, for reasons that we detail in this paper, the direct extension of the established economic MPC dissipativity theory to MDPs is restricted to some very specific problems, see, e.g., [20].

As an alternative to a direct extension of the existing theory, we propose in this paper to form a generalization of the economic MPC dissipativity theory built on the measure space underlying the MDP rather than on the state space itself. This approach yields stronger results than stability to a set, and is not restricted to a specific class of stochastic dynamics. For the sake of clarity, we ought to stress here that the dissipativity theory we will develop is to be understood as a generalization of the dissipativity theory for economic MPC, rather than as a general dissipativity theory for stochastic systems investigated in, e.g., [17,11].

Contributions: In this paper we propose a generalization of both the established dissipativity theory and the stochastic dissipativity theory of [20], and provide a stronger discussion on the stability of MDPs than convergence to a set, and a more specific stability theory than the generic discussions on the stability of Markov Chains. We show that the philosophy underlying the established dissipativity theory for economic MPC is valid in the MDP context, but its application requires generalizing the concept of stage cost and storage functions to functionals operating on the set of probability measures (or densities). The classic notion of norm must then be replaced by the notion of dissimilarity measures such as, e.g., the Kullback-Leibler divergence, the Wasserstein metric, or the total variational distance [14]. Strong stability results follow, where dissipativity ensures that the measures underlying the MDP converge asymptotically to the steady-state optimal measure. This generalization is illustrated in the Linear Quadratic Regulator (LQR) case subject to a Gaussian process noise, for which an explicit storage functional is provided.

The paper is organized as follows. Section 2 proposes a discussion of the difficulties of extending the established dissipativity theory to MDPs. Section 3 proposes a generalization of the classical Lyapunov-based stability arguments for optimal policies to MDPs, using a functional approach. The resulting stability theory shows that the cost function normally used in MDPs, made of the expected sum of stage costs, does not satisfy the necessary criterion for stability, and that a cost rotation is in general needed. Section 4 generalizes the concept of rotation, dissipativity and storage function, allowing for a discussion of MDP stability in the Lyapunov context. In section 5, these concepts are deployed in the stochastic LQR context, showing that the proposed approach is sensible. Section 6 concludes the paper.

2 Markov Decision Processes and Dissipativity Theory

In this section, we detail the Markov Decision Processes considered in this paper, and provide a brief introduction to the state of the art on dissipativity for Economic MPC, providing a framework to discuss the closed-loop stability of optimal control policies. We then discuss and detail formally why the established dissipativity theory cannot readily apply to stochastic problems, see Lemma 1 and Remark 2, and hence motivate the extension of the established dissipativity theory to MDPs, which is provided in this paper.

We consider discrete dynamical systems evolving on a continuous state space over $\mathbb{R}^n$, with stochastic states $\boldsymbol{s}_k \in \mathbb{R}^n$, where k denotes the discrete time. The underlying measure space for $\boldsymbol{s}_k$ is $\mathbb{R}^n$ equipped with the Lebesgue measure υ as reference measure, and the set of Lebesgue-measurable sets as σ-algebra $\mathcal{S}$. The actions (control inputs) $\boldsymbol{a}_k$ are taken in the continuous space $\mathbb{R}^m$. We then consider stochastic dynamics defined by the conditional probability measure ξ:

$$\xi\left[\mathcal{B}\,|\,\boldsymbol{s}_k,\boldsymbol{a}_k\right],\quad \boldsymbol{s}_{k+1}\in\mathcal{B}, \tag{1}$$

defining the conditional probability of observing a transition from a given state-action pair $\boldsymbol{s}_k, \boldsymbol{a}_k$ to a subsequent state $\boldsymbol{s}_{k+1}$ in the Lebesgue-measurable set $\mathcal{B} \subseteq \mathbb{R}^n$. Furthermore, we consider deterministic causal policies $\boldsymbol{\pi} : \mathbb{R}^n \to \mathbb{R}^m$ such that:

$$\boldsymbol{a}_k = \boldsymbol{\pi}\left(\boldsymbol{s}_k\right). \tag{2}$$

We label $\mathcal{M}$ the set of probability measures over $\mathbb{R}^n$ and Π the set of policies, i.e., $\boldsymbol{\pi} \in \Pi$. A policy $\boldsymbol{\pi}$ in closed loop with dynamics (1) generates a closed-loop Markov Chain with the underlying sequence of probability measures $\rho_k \in \mathcal{M},\ k = 0, ..., \infty$ describing the stochasticity of the MDP states $\boldsymbol{s}_k$. Hence, in the following, for each $k = 0, \ldots, \infty$, $\boldsymbol{s}_k$ will be a sequence of random variables, and ρ_k the associated sequence of probability measures, i.e., $\boldsymbol{s}_k \sim \rho_k$. We then define the transition operator $\mathcal{T}_{\boldsymbol{\pi}} : \mathcal{M} \times \Pi \to \mathcal{M}$ as the map from a measure ρ_k to its successor ρ_{k+1} via (1) and under policy $\boldsymbol{\pi}$. More specifically, $\mathcal{T}_{\boldsymbol{\pi}}$ is formally defined as [14]

$$\rho_{k+1}(\cdot) = \mathcal{T}_{\boldsymbol{\pi}}\,\rho_k(\cdot) = \int \xi\left[\,\cdot\,|\,\boldsymbol{s},\boldsymbol{\pi}\left(\boldsymbol{s}\right)\right]\rho_k\left(\mathrm{d}\boldsymbol{s}\right). \tag{3}$$

In the following, we will restrict our policies $\boldsymbol{\pi}$ to be in the set $P \subseteq \Pi$ such that the integral in (3) is well-defined for all Borel sets. The triple $(S_{\boldsymbol{s}}, \mathcal{S}_{\boldsymbol{s}}, \mathbb{P}_{\boldsymbol{\pi}})$ defines the probability space associated with a Markov chain, where $S_{\boldsymbol{s}} = \prod_{k=0}^{\infty} \mathbb{R}^n$, with associated σ-field $\mathcal{S}_{\boldsymbol{s}}$, and $\mathbb{P}_{\boldsymbol{\pi}}$ is the probability measure defined by (1)-(2) [14].

Note that for most of the discussions in this paper, the sequence of probability measures $\rho_{0,\ldots,\infty}$ can also be interpreted as a sequence of probability densities if the measures $\rho_{0,\ldots,\infty}$ have a Radon-Nykodim derivative with respect to the Lebesgue measure. In that case, we will assume that the associated probability densities are all in $\mathcal{L}^p\left(\mathbb{R}^n, \mathcal{S}, \upsilon\right)$. The use of measures instead of densities here is a technicality aimed at providing a generic discussion. Let us label $\mathbb{E}_{\boldsymbol{s}\sim\rho_k}[\cdot]$ the expected value operator with respect to probability measure $\rho_k \in \mathcal{M}$. Furthermore, let us define the stage cost function $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. We then consider undiscounted MDPs [16] over an infinite horizon, of the classic form:

$$\begin{aligned} J_\star\left[\rho\right] = \min_{\boldsymbol{\pi}\in P} \quad & \sum_{k=0}^{\infty} \mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L\left(\boldsymbol{s},\boldsymbol{\pi}\left(\boldsymbol{s}\right)\right) - L_0\right] && \text{(4a)} \\ \text{s.t.} \quad & \rho_{k+1} = \mathcal{T}_{\boldsymbol{\pi}}\,\rho_k, \quad \rho_0 = \rho, && \text{(4b)} \end{aligned}$$

where $J_\star : \mathcal{M} \to \mathbb{R}$, and the argument of $J_\star$, i.e., $\rho \in \mathcal{M}$, specifies the initial condition of the Markov Chain (4b). Constant $L_0 \in \mathbb{R}$ is the optimal cost of the optimal steady-state problem

$$\begin{aligned} L_0 = \min_{\boldsymbol{\pi}\in P,\,\rho} \quad & \mathbb{E}_{\boldsymbol{s}\sim\rho}\left[L\left(\boldsymbol{s},\boldsymbol{\pi}\left(\boldsymbol{s}\right)\right)\right] && \text{(5a)} \\ \text{s.t.} \quad & \rho = \mathcal{T}_{\boldsymbol{\pi}}\,\rho, && \text{(5b)} \end{aligned}$$

which also delivers the corresponding optimal steady-state measure $\rho_\star \in \mathcal{M}$. We assume that $\rho_\star$ exists and is unique. Problem (4) defines an optimal policy $\boldsymbol{\pi}_\star : \mathbb{R}^n \to \mathbb{R}^m$ in P.

For simplicity, we assume throughout the paper that Problem (4) has a unique minimizer. We do not expect the theory presented hereafter to change if $\boldsymbol{\pi}$ solution of (4) is not unique or defined via an infimization rather than a minimization problem. This would arguably require additional technicalities, though, which we avoid here for the sake of simplicity.

Remark 1 *In order to frame Problem* (4) *in the general context of undiscounted MDPs, we observe that, in case L_0 is also the optimal average cost, given by*

$$\begin{aligned} L_0 = \min_{\boldsymbol{\pi}\in P} \quad & \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L\left(\boldsymbol{s},\boldsymbol{\pi}\left(\boldsymbol{s}\right)\right)\right] && \text{(6a)} \\ \text{s.t.} \quad & \rho_{k+1} = \mathcal{T}_{\boldsymbol{\pi}}\,\rho_k, \quad \rho_0 = \rho, && \text{(6b)} \end{aligned}$$

$\forall\,\rho$, then Problem (4) *yields bias optimality, a formal definition of which is available in, e.g., [16]. This observation is only provided to give additional context and does not impact the remainder of this paper.*

We ought to observe that, while in this paper we focus on undiscounted MDPs of the form (4), we expect that our framework can be readily extended to cover the discounted case by building on the ideas proposed in [22]. This extension is the subject of current research.

In this paper, we are interested in characterizing conditions on the MDP dynamics and stage cost L such that the optimal policy $\boldsymbol{\pi}_\star$ solution of (4) is stabilizing the closed-loop Markov Chain to the optimal steady-state solution of (5), i.e., such that

$$\lim_{k\to\infty} \rho_k = \rho_\star, \tag{7}$$

in some sense that we will discuss.

We recall that in the special case where (4) is deterministic, such that ρ_k reduces to a sequence of Dirac measures, the stability of (4) can be discussed in the framework of the established dissipativity theory for economic MPC, using the concept of storage function. In that context, a storage function $\boldsymbol{\lambda} : \mathbb{R}^n \to \mathbb{R}$ is sought, such that

$$L\left(\boldsymbol{s}_k, \boldsymbol{\pi}\left(\boldsymbol{s}_k\right)\right) - \boldsymbol{\lambda}\left(\boldsymbol{s}_{k+1}\right) + \boldsymbol{\lambda}\left(\boldsymbol{s}_k\right) \geq \varrho\left(\|\boldsymbol{s}_k - \boldsymbol{s}_\star\|\right) \tag{8}$$

holds over the deterministic system trajectories for an optimal steady state $\boldsymbol{s}_\star$ of the system and a class-$\mathcal{K}_\infty$ [1] function $\varrho : \mathbb{R}_+ \to \mathbb{R}_+$. Under the condition that the storage function remains bounded over the prediction horizon, the optimal value function $J_\star^{\mathrm{R}}$ resulting from the *rotated* cost $L^{\mathrm{R}} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$ defined as

$$L^{\mathrm{R}}\left(\boldsymbol{s}_k, \boldsymbol{\pi}\left(\boldsymbol{s}_k\right)\right) = L\left(\boldsymbol{s}_k, \boldsymbol{\pi}\left(\boldsymbol{s}_k\right)\right) - \boldsymbol{\lambda}\left(\boldsymbol{s}_{k+1}\right) + \boldsymbol{\lambda}\left(\boldsymbol{s}_k\right) \tag{9}$$

is a Lyapunov function for the system. The general philosophy of the dissipativity theory is to transform the stage cost L of an economic optimal control problem (4) into a new stage cost L^{R} that yields the same optimal policy, while resulting in a value function $J_\star^{\mathrm{R}}$ that is a Lyapunov function for the closed-loop trajectories.

A direct extension of this philosophy to treat the stability of MDPs in the form (4) is appealing. This approach has been followed in [20] for Markovian switching systems, where the rotated cost is given by

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L^{\mathrm{R}}\left(\boldsymbol{s}, \boldsymbol{\pi}\left(\boldsymbol{s}\right)\right)\right] = \\ &\mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L\left(\boldsymbol{s}, \boldsymbol{\pi}\left(\boldsymbol{s}\right)\right)\right] + \mathbb{E}_{\boldsymbol{s}\sim\rho_k, \boldsymbol{s}_+\sim\rho_{k+1}}\left[\boldsymbol{\lambda}\left(\boldsymbol{s}_+\right) - \boldsymbol{\lambda}\left(\boldsymbol{s}\right)\right].\end{aligned} \tag{10}$$

and the stage cost $\mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L^{\mathrm{R}} - L_0\right]$ is used in (4). Note that if (5) is formulated by replacing L with L^{R}, its optimal cost is still L_0. In order to prove that $J_\star^{\mathrm{R}}$ is non-increasing using the standard approach, one then needs

$$\mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L^{\mathrm{R}}\left(\boldsymbol{s}, \boldsymbol{\pi}\left(\boldsymbol{s}\right)\right) - L_0\right] \geq \mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[\varrho\left(\|\boldsymbol{s} - \boldsymbol{s}_\star\|\right)\right] \tag{11}$$

to hold along the system trajectories. This condition has been called *strict stochastic dissipativity* in [20]. However, except for some special cases (e.g., Markovian switching systems), $L^{\mathrm{R}} - L_0$ cannot be non-negative everywhere, such that by construction (11) cannot hold. This statement is supported in a more formal way by the following lemma. Note that the lemma discusses the case of a rotated cost L^{R}, but also applies to the original MDP (4) if one selects $\boldsymbol{\lambda}(\boldsymbol{s}) = 0$, such that $L^{\mathrm{R}} = L$.

Lemma 1 *For infinite-horizon, undiscounted MDPs on the continuous state space $\mathbb{R}^n$, the following statements cannot be all simultaneously true:*

1. *$L^{\mathrm{R}}\left(\boldsymbol{s}, \boldsymbol{\pi}\left(\boldsymbol{s}\right)\right) - L_0 = 0$ on a set $S_0 \subset \mathbb{R}^n$ of zero Lebesgue measure and $L^{\mathrm{R}}\left(\boldsymbol{s}, \boldsymbol{\pi}\left(\boldsymbol{s}\right)\right) - L_0 \geq \varrho\left(\|\boldsymbol{s}\|_{S_0}\right)$ holds for all $\boldsymbol{s} \in \mathbb{R}^n$, where ϱ is a continuous class $\mathcal{K}_\infty$ function, and $\|\cdot\|_{S_0}$ a continuous distance to S_0.*
2. *$J_\star^R[\rho_0]$ exists and is bounded for a non-empty set of initial probability measures ρ_0.*
3. *There is a $k_0 \in \mathbb{N}_+$ and a constant $0 < b < \infty$ such that for every $k \geq k_0$ the measures ρ_k are equipped with probability density functions $f_k : \mathbb{R}^n \to \mathbb{R}_+$ such that $f_k(\boldsymbol{s}) \leq b$, $\forall\, \boldsymbol{s} \in \mathbb{R}^n$.*

In order to keep the proof accessible, the argument is developed using calculus, making it fairly long but simple.

PROOF. By contradiction. In order for $J_\star^{\mathrm{R}}[\rho_0]$ to exist and be bounded (statement 2.), the limit

$$\lim_{k\to\infty} \mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L^{\mathrm{R}}\left(\boldsymbol{s}, \boldsymbol{\pi}\left(\boldsymbol{s}\right)\right) - L_0\right] = 0 \tag{12}$$

must hold. In order to prove the contradiction, let us assume that statements 1. and 3. hold. Let us define the sub-level set $S_\alpha \subseteq \mathbb{R}^n$ as:

$$S_\alpha = \left\{\, \boldsymbol{s} \in \mathbb{R}^n \quad \text{s.t.} \quad \varrho\left(\|\boldsymbol{s}\|_{S_0}\right) \leq \alpha \right\}. \tag{13}$$

Because ϱ and $\|\cdot\|_{S_0}$ are continuous, and $\varrho\left(\|\boldsymbol{s}\|_{S_0}\right) = 0$ for some $\boldsymbol{s}$, S_α has a non-zero Lebesgue measure for any $\alpha > 0$. Let us further define the set $\mathcal{S} \subset \mathbb{R}^n$ as:

$$\mathcal{S} = S_{\bar\alpha} \quad \text{where} \quad \bar\alpha > 0 \quad \text{is s.t.} \quad \int_{S_{\bar\alpha}} \mathrm{d}\boldsymbol{s} = b^{-1}, \tag{14}$$

whose existence is guaranteed by the continuity of ϱ and $\|\cdot\|_{S_0}$. Let us then define the probability density f_b as

$$f_b\left(\boldsymbol{s}\right) = \begin{cases} b \text{ if } \boldsymbol{s} \in \mathcal{S} \\ 0 \text{ if } \boldsymbol{s} \notin \mathcal{S} \end{cases}. \tag{15}$$

Note that f_b is indeed a probability density, since by construction $\int_{\mathbb{R}^n} f_b \mathrm{d}\boldsymbol{s} = 1$. We will show next that density f_b yields a strictly positive lower bound for

[1] We define $\mathbb{R}_+ := \{\, x \in \mathbb{R} \mid x \geq 0 \,\}$. Function $\varrho : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies $\varrho \in \mathcal{K}$ if it is continuous, zero at zero and strictly increasing. If additionally $\varrho \in \mathcal{K}$ is radially unbounded, then $\varrho \in \mathcal{K}_\infty$.

$\mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L^{\mathrm{R}}\left(\boldsymbol{s},\boldsymbol{\pi}\left(\boldsymbol{s}\right)\right)-L_0\right]$. We first observe that since S_0 is of zero Lebesgue measure, $\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)>0$ almost everywhere in $\mathcal{S}$, and $\mathcal{S}$ has a strictly positive Lebesgue measure, there is a constant c such that

$$\int_{\mathbb{R}^n}\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)f_b\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}=b\int_{\mathcal{S}}\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)\mathrm{d}\boldsymbol{s}=c>0 \tag{16}$$

holds. Let us define $\Delta\left(\boldsymbol{s}\right)=f_k\left(\boldsymbol{s}\right)-f_b\left(\boldsymbol{s}\right)$. We observe that since f_k and f_b are both probability densities, equality $\int_{\mathbb{R}^n}\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}=0$ holds, such that

$$\int_{\mathcal{S}}\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}=-\int_{\mathcal{S}^{\mathrm{c}}}\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}\leq 0, \tag{17}$$

where $\mathcal{S}^{\mathrm{c}}$ is the complementary set to $\mathcal{S}$, and where the inequality holds because for all $\boldsymbol{s}\in\mathcal{S}$

$$\Delta\left(\boldsymbol{s}\right)=\underbrace{f_k\left(\boldsymbol{s}\right)}_{\leq b}-\underbrace{f_b\left(\boldsymbol{s}\right)}_{=b}\leq 0. \tag{18}$$

Furthermore, from the definition of $\mathcal{S}$, using (13) we have that:

$$0\leq\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)\leq\bar{\alpha}\qquad\forall\boldsymbol{s}\in\mathcal{S}, \tag{19a}$$
$$\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)>\bar{\alpha}\qquad\forall\boldsymbol{s}\in\mathcal{S}^{\mathrm{c}}. \tag{19b}$$

Using (19) and (17), we then observe that

$$0\geq\int_{\mathcal{S}}\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}\geq\bar{\alpha}\int_{\mathcal{S}}\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}, \tag{20a}$$
$$\begin{aligned}\int_{\mathcal{S}^{\mathrm{c}}}\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}&\geq\bar{\alpha}\int_{\mathcal{S}^{\mathrm{c}}}\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}\\&=-\bar{\alpha}\int_{\mathcal{S}}\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}\geq 0\end{aligned} \tag{20b}$$

Hence by summing (20a)-(20b) we observe that:

$$\int_{\mathbb{R}^n}\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)\Delta\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}\geq 0. \tag{21}$$

Using (21), (16) and the definition of Δ yields:

$$\int_{\mathbb{R}^n}\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)f_k\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}\geq\int_{\mathbb{R}^n}\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)f_b\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}=c. \tag{22}$$

We can finally conclude by observing that for all $k\geq k_0$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L^{\mathrm{R}}\left(\boldsymbol{s},\boldsymbol{\pi}\left(\boldsymbol{s}\right)\right)-L_0\right]&\geq\int_{\mathbb{R}^n}\varrho\left(\|\boldsymbol{s}\|_{S_0}\right)f_k\left(\boldsymbol{s}\right)\mathrm{d}\boldsymbol{s}\\&\geq c>0.\end{aligned} \tag{23}$$

This last inequality is in contradiction with (12), such that statements 1., 2. and 3. cannot hold together. □

Remark 2 *The consequence of Lemma 1 is that the stage cost L of an MDP can be rotated such that the resulting value function is a Lyapunov function in some very specific cases only, i.e.:*

- *The MDP state $\boldsymbol{s}_k$ converges to a set of zero Lebesgue measure in $\mathbb{R}^n$. If the sequence of measures ρ_k, $k=0,\ldots,\infty$ admits probability densities, these densities are unbounded as $k\rightarrow\infty$. If, e.g., S_0 is a point in $\mathbb{R}^n$, the limit of the sequence ρ_k, $k=0,\ldots,\infty$ is a Dirac measure. This requires the MDP dynamics to have very specific properties, such as, e.g., vanishing perturbations. Another instance in which this situation can occur is if (1) represents a Markovian switching system, see [20].*
- *The cost rotation makes $L^{\mathrm{R}}\left(\boldsymbol{s},\boldsymbol{\pi}_\star(\boldsymbol{s})\right)-L_0=0$ on a measurable set $S_0\subseteq\mathbb{R}^n$. The stability discussion is then limited to stating that the sequence of probability measures ρ_k, with $k=0,\ldots,\infty$ will converge to a probability measure ρ_∞ that has its entire support in S_0, i.e., $\rho_\infty(S_0)=1$. Concretely, the MDP state $\boldsymbol{s}_k$ will then converge to S_0 with probability 1, but nothing can be said of the evolution of $\boldsymbol{s}_k$ within S_0. Such situations are typical of robust MPC, as studied in, e.g., [2,3].*

A direct extension of the established dissipativity theory cannot apply to a more general context.

In this paper, we will show that the issues detailed above do not stem from the philosophy underlying the established dissipativity theory, but simply from considering a cost rotation (10) that is restricted to being linear in the probability measures ρ_k.

3 Lyapunov Stability for MDPs

The previous section details why the established dissipativity theory does not readily apply to stochastic problems. In this section, we will show that the issue can be addressed within the general philosophy of the established dissipativity theory, but that it requires an extension where the classic state-space of deterministic systems must be replaced by the set of probability measures underlying the stochastic systems. In order to develop this extension, let us generalize MDP (4) into the following optimal control problem:

$$V_\star\left[\rho\right]=\min_{\boldsymbol{\pi}\in P}\quad\sum_{k=0}^{\infty}\mathcal{L}\left[\rho_k,\boldsymbol{\pi}\right] \tag{24a}$$
$$\text{s.t.}\quad\rho_{k+1}=\mathcal{T}_{\boldsymbol{\pi}}\,\rho_k,\quad\rho_0=\rho, \tag{24b}$$

where $V_\star:\mathcal{M}\rightarrow\mathbb{R}$, and the argument of $V_\star$, i.e., ρ, specifies the initial condition of the Markov Chain (24b). The stage cost $\mathcal{L}\in\mathcal{M}\times P\rightarrow\mathbb{R}$ is a (possibly nonlinear) functional over the probability measures ρ_k, and the policy $\boldsymbol{\pi}$. We assume here that $V_\star\left[\rho\right]$ is finite for a non-empty set of measures ρ. One can readily observe that Problem (4) is a special case of (24), obtained by selecting

$$\mathcal{L}\left[\rho_k, \boldsymbol{\pi}\right] = \mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L\left(\boldsymbol{s}, \boldsymbol{\pi}\left(\boldsymbol{s}\right)\right) - L_0\right]. \tag{25}$$

Note that the condition $\mathcal{L}\left[\rho_\star, \boldsymbol{\pi}_\star\right] = 0$ is required for $V_\star$ to be bounded. In the specific case of Problem (4), this can further be seen by using the definition of Equation (5) in (25). We will see in this paper that the freedom of using more general functionals than (25) for $\mathcal{L}$ is the key to generalizing the economic NMPC dissipativity theory to MDPs.

We detail next how (24) makes it possible to build a fairly straightforward and classic Lyapunov stability result on the set of probability measures. To that end, let us introduce the following key concepts.

Definition 1 (Dissimilarity measure) *Let us define a dissimilarity measure $D\left(\cdot\,||\cdot\right) : \mathcal{R}\times\mathcal{R}\rightarrow\mathbb{R}$ as an application from a subset of probability measures to the real positive numbers, such that:*

$$D\left(\rho\,||\,\rho'\right) \geq 0 \quad \textit{and} \quad D\left(\rho\,||\,\rho\right) = 0, \quad \forall\,\rho,\rho'\in\mathcal{R}, \tag{26}$$

where $\mathcal{R}\subseteq\mathcal{M}$ is (a subset of) the set of probability measures.

A useful example of dissimilarity measures is the Kullback-Leibler divergence (D_{KL}) defined as

$$D_{\mathrm{KL}}\left(\rho\,||\,\rho'\right) = \int\log\frac{\mathrm{d}\rho}{\mathrm{d}\rho'}\mathrm{d}\rho = \int f(\boldsymbol{s})\log\frac{f(\boldsymbol{s})}{f'(\boldsymbol{s})}\mathrm{d}\boldsymbol{s}, \tag{27}$$

where $\frac{\mathrm{d}\rho}{\mathrm{d}\rho'}$ is the Radon-Nykodim derivative of ρ with respect to ρ' and the second equality holds if ρ, ρ' have underlying probability densities f, f'. Other useful examples in control are the Wasserstein metric, and the total variational distance [14]. The notion of stability on the set of probability measures can then be formalized as follows.

Definition 2 (D-Stability) *A Markov Chain is D-stable with respect to probability measure $\rho_\star$ and dissimilarity measure D if, for any $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that $D\left(\rho_0\,||\,\rho_\star\right) < \delta$ implies $D\left(\rho_k\,||\,\rho_\star\right) < \epsilon$ for all $k \geq 0$. If, moreover, the probability measure $\rho_\star$ is D-attractive, i.e.,*

$$\lim_{k\rightarrow\infty} D\left(\rho_k\,||\,\rho_\star\right) = 0, \tag{28}$$

holds, then the Markov Chain is D-asymptotically stable.

□

Note that D-stability is introduced as a practical way to encompass many commonly used stability concepts. If, e.g., D is the total variational distance, then the obtained stability is often referred to as ergodicity [14]. If, e.g., D is the expected value under ρ of the square of $\boldsymbol{s}$, then one obtains asymptotic stability for deterministic systems [19] and mean square stability for stochastic systems [20].

The next theorem formalizes the stability of OCP (24) on the set of probability measures, following the same arguments as classic Lyapunov stability for optimal policies over deterministic problems.

Theorem 1 *Assume that the inequalities*

$$\mathcal{L}\left[\rho_k, \boldsymbol{\pi}_\star\right] \geq \varrho_1\left(D\left(\rho_k\,||\,\rho_\star\right)\right), \tag{29a}$$

$$V_\star[\rho_k] \leq \varrho_2\left(D\left(\rho_k\,||\,\rho_\star\right)\right), \tag{29b}$$

hold for some class-$\mathcal{K}_\infty$ functions $\varrho_{1,2}$ and for all $\rho\in\Xi\subseteq\mathcal{R}$, where set Ξ is a non-empty set such that $V_\star < \infty$ on Ξ. Then the Markov chain is D-asymptotically stable with respect to the probability measure $\rho_\star$ and dissimilarity measure D for any $\rho_0\in\Xi$.

□

Remark 3 *Note that assumption* (29b) *corresponds to a standard assumption in the context of MPC—referred to as a form of weak controllability [19]—which requires that the value function is upper-bounded by a class-$\mathcal{K}_\infty$ function of a norm of $\boldsymbol{s}-\boldsymbol{s}_\star$.*

Remark 4 *Condition* (29a) *can be mistaken for the simple requirement that the cost functional $\mathcal{L}$ should correspond to a stochastic tracking MPC scheme. This interpretation is, however, not necessarily correct. Indeed, a stochastic tracking MPC scheme would typically use a cost functional of the form*

$$\mathcal{L}\left[\rho_k, \boldsymbol{\pi}_\star\right] = \frac{1}{2}\mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[\boldsymbol{s}^\top Q\boldsymbol{s} + \boldsymbol{\pi}_\star\left(\boldsymbol{s}\right)^\top R\boldsymbol{\pi}_\star\left(\boldsymbol{s}\right)\right], \tag{30}$$

which typically does not satisfy condition (29a) *unless $\rho_\star$ is a Dirac measure centered at $\boldsymbol{s}=0$. The latter requires that some very specific properties are satisfied by the system dynamics* (1)*, and hence does not hold in general. See Lemma 1 and the following remarks.*

PROOF. (of Theorem 1) We first observe that because $V_\star$ is bounded on Ξ, Ξ is positive invariant for system ξ defined in (1) under policy $\boldsymbol{\pi}_\star$ solving (24) and

$$V_\star[\rho_{k+1}] - V_\star[\rho_k] = -\mathcal{L}\left[\rho_k, \boldsymbol{\pi}_\star\right] \leq -\varrho_1\left(D\left(\rho_k\,||\,\rho_\star\right)\right) \tag{31}$$

holds on Ξ. Furthermore, we observe that from (29a), the bound

$$V_\star[\rho_k] \geq \mathcal{L}\left[\rho_k, \boldsymbol{\pi}_\star\right] \geq \varrho_1\left(D\left(\rho_k \,||\, \rho_\star\right)\right) \geq 0 \tag{32}$$

holds for any $\rho_k \in \Xi$. Hence $V_\star[\rho_k] \geq 0$ is bounded and monotonically decreasing on Ξ, such that it must converge to a finite positive value $\bar{V}$ as $k \to \infty$. We then need to prove that $\bar{V} = 0$. To that end, consider $\delta, \epsilon > 0$ selected as

$$D\left(\rho_0 \,||\, \rho_\star\right) \leq \delta, \qquad \epsilon = \varrho_1^{-1}(\varrho_2(\delta)), \tag{33}$$

such that

$$V_\star[\rho_0] \leq \varrho_2(\delta) = \varrho_1(\epsilon). \tag{34}$$

Using (31) and (32), we observe that for all k:

$$\begin{aligned} D\left(\rho_k \,||\, \rho_\star\right) &\leq \varrho_1^{-1}(V_\star[\rho_k]) \leq \varrho_1^{-1}(V_\star[\rho_0]) \\ &\leq \varrho_1^{-1}(\varrho_1(\epsilon)) = \epsilon, \end{aligned} \tag{35}$$

which proves stability. In order to prove attractivity, we proceed by contradiction. Assume that

$$\lim_{k\to\infty} V_\star[\rho_k] = \bar{V} > 0, \tag{36}$$

then using (29b) and (35), the inequalities

$$\varrho_2^{-1}(\bar{V}) \leq \lim_{k\to\infty} D\left(\rho_k \,||\, \rho_\star\right) \leq \varrho_1^{-1}(\bar{V}) \tag{37}$$

hold. Using (31) we obtain:

$$V_\star[\rho_k] \leq V_\star[\rho_0] - \sum_{j=0}^{k} \varrho_1\left(D\left(\rho_j \,||\, \rho_\star\right)\right). \tag{38}$$

Since $D\left(\rho_k \,||\, \rho_\star\right)$ converges to the interval $[\varrho_2^{-1}(\bar{V}), \varrho_1^{-1}(\bar{V})]$, then

$$\lim_{k\to\infty} \varrho_1\left(D\left(\rho_k \,||\, \rho_\star\right)\right) \geq \varrho_1\left(\varrho_2^{-1}(\bar{V})\right) > 0, \tag{39}$$

such that $V_\star[\rho_k] \to -\infty$ as $k \to \infty$, which is in contradiction with (32). Consequently, $\bar{V} = 0$, and (28) must hold. □

Note that the stability result of this theorem can carry several meanings, depending on the properties of D. If, e.g., $D(\rho_1||\rho_2) = \left\| \mathbb{E}_{\boldsymbol{s}\sim\rho_1}\left[\boldsymbol{s}\right] - \mathbb{E}_{\boldsymbol{s}\sim\rho_2}\left[\boldsymbol{s}\right] \right\|$, then only the expected value of the state is guaranteed to converge. In case the selected dissimilarity measure carries stronger properties, stronger stability results arise. Let us detail a useful special case in the next corollary.

Corollary 1 *Assume that the assumptions of Theorem 1 hold, and the dissimilarity measure $D\left(\rho \,||\, \rho_\star\right)$ is such that $D\left(\rho \,||\, \rho_\star\right) = 0$ implies that $\rho = \rho_\star$ almost everywhere. Then*

$$\lim_{k\to\infty} \rho_k\left(\cdot\right) = \rho_\star\left(\cdot\right) \tag{40}$$

holds almost everywhere.

□

PROOF. The limit (28) follows from Theorem 1. By the properties assumed on the dissimilarity measure, this directly entails (40). □

Examples of dissimilarity measure satisfying Corollary 1 include D_{KL} and the total variational distance. It may be useful here to discuss what form of stability is established in Theorem 1. Stability proofs in the context of Classic MPC and Economic MPC discuss the behavior of single trajectories, starting from arbitrary initial conditions in a set, and proves the convergence to an optimal steady-state. In Robust MPC one discusses the behavior of all possible stochastic trajectories, and proves the convergence to a set, without describing the behavior inside that set. In contrast, Theorem 1 discusses the behavior of trajectories by showing that their asymptotic behavior is to be distributed according to a distribution with zero dissimilarity with respect to the optimal steady-state measure of the MDP. For suitably selected dissimilarity measures (see, e.g., Corollary 1), this entails that these two distributions must coincide almost everywhere.

We now turn to discussing how the stability of the MDP resulting from a generic stage cost functional $\mathcal{L}$ can be discussed in terms of (29a) via functional cost rotations.

4 Functional Cost Rotations

Making a Lyapunov stability argument on problem (24) requires the cost functional $\mathcal{L}\left[\rho_k, \boldsymbol{\pi}\right]$ to satisfy (29a). Following the arguments of Lemma 1, one can readily observe that, in general, for MDP (4) to be well-posed, the MDP stage cost $L - L_0$ cannot be strictly positive. As a result, when recasting a given MDP (4) in its equivalent functional form (24) using identity (25), the resulting functional stage cost $\mathcal{L}\left[\rho, \boldsymbol{\pi}\right]$ cannot be positive for all probability measures ρ, such that (29a) cannot hold. This challenges by construction the extension of classical Lyapunov stability to general MDPs. As a result, a classic rotation in the form (10) is not applicable in general.

Fortunately, it is possible to tackle these difficulties by adopting a more general cost rotation than (10). More specifically, we will consider functional cost rotations of the form:

$$\mathcal{L}^{\mathrm{R}}\left[\rho_k, \boldsymbol{\pi}\right] = \mathcal{L}\left[\rho_k, \boldsymbol{\pi}\right] - \lambda\left[\rho_{k+1}\right] + \lambda\left[\rho_k\right], \tag{41}$$

where $\lambda : \mathcal{M} \to \mathbb{R}$ is a (possibly) nonlinear functional. Rotation (10) is then a special case of (41), where the form

$$\lambda\left[\rho_k\right] = \mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[\lambda\left(\boldsymbol{s}\right)\right] \tag{42}$$

is imposed. Following the arguments presented above, the form (42) is in general not able to deliver a Lyapunov function.

Similarly to classical cost rotations, we observe that (41) leaves the policy solution of (24) unchanged, as long as $\lambda[\rho_k]$ is bounded for all k. A generalized dissipativity criterion can then be formulated as follows.

Definition 3 (Functional Strict Dissipativity) *There exists a functional $\lambda : \mathcal{M} \to \mathbb{R}$ and a class-$\mathcal{K}_\infty$ function ϱ such that $\mathcal{L}^{\mathrm{R}}\left[\rho_k, \boldsymbol{\pi}\right]$ defined by* (41) *satisfies* (29a)*, i.e.,*

$$\mathcal{L}\left[\rho_k, \boldsymbol{\pi}\right] - \lambda\left[\rho_{k+1}\right] + \lambda\left[\rho_k\right] \geq \varrho\left(D\left(\rho_k \,||\, \rho_\star\right)\right) \tag{43}$$

holds for all $\rho_k \in \mathcal{R}$ such that $V_\star\left(\rho_k\right)$ is finite.

As we will prove next, the functional dissipativity criterion (43) then yields D-asymptotic stability. Indeed, let us define a rotated problem as:

$$V_\star^{\mathrm{R}}\left[\rho\right] = \min_{\boldsymbol{\pi}\in P} \quad \lim_{N\to\infty} \sum_{k=0}^{N-1} \mathcal{L}^{\mathrm{R}}\left[\rho_k, \boldsymbol{\pi}\right] + \lambda\left[\rho_N\right] \tag{44a}$$

$$\text{s.t.} \quad \rho_{k+1} = \mathcal{T}_{\boldsymbol{\pi}}\, \rho_k, \quad \rho_0 = \rho, \tag{44b}$$

where $V_\star^{\mathrm{R}} : \mathcal{M} \to \mathbb{R}$, and the argument of $V_\star^{\mathbb{R}}$, i.e., ρ, specifies the initial condition of the Markov Chain (44b). We then establish D-asymptotic stability in the next theorem. Under the assumption that functional strict dissipativity holds for a bounded storage functional λ, the existence of the limit in (44a) will become clear in equation (46) in the proof of the next theorem.

Theorem 2 *Assume that there exists a storage functional λ bounded from above and below, and satisfying* (43)*. Assume moreover that*

$$V_\star^{\mathrm{R}}[\rho_k] \leq \varrho_2(D\left(\rho_k \,||\, \rho_\star\right)). \tag{45}$$

with $V_\star^{\mathrm{R}} : \mathcal{M} \to \mathbb{R}$ defined in (44)*. Then, the rotated problem* (44) *and the original problem* (24) *deliver the same optimal policy. Moreover, the Markov chain is D-asymptotically stable with respect to the probability measure $\rho_\star$ and dissimilarity measure D.*

Remark 5 *We observe that* (45) *can be interpreted as a controllability assumption, since it holds whenever ρ_k can be steered to $\rho_\star$ (in the sense of the dissimilarity measure D) with finite cost. This is equivalent, mutatis mutandis, to the deterministic case, see Remark 3.*

Remark 6 *A reader well acquainted to the dissipativity theory for economic MPC will recognize in Theorem 2 a generalization of the theory applicable in the deterministic case, where a criterion similar to* (43) *and a bound similar to* (45) *entail the convergence of the system state to the optimal steady-state. Theorem 2 extends these concepts to the convergence of the system in the sense of the probability measures underlying the dynamics rather than in the sense of the states themselves.*

PROOF. (of Theorem 2) The first claim follows from standard arguments, since boundedness of $\lambda[\rho_k]$ entails

$$\sum_{k=0}^{N-1} \mathcal{L}^{\mathrm{R}}\left[\rho_k, \boldsymbol{\pi}\right] + \lambda\left[\rho_N\right] = \sum_{k=0}^{N-1} \mathcal{L}\left[\rho_k, \boldsymbol{\pi}\right] + \lambda\left[\rho_0\right]. \tag{46}$$

By taking the limit $N \to \infty$, (which exists if the original problem (24) is well-posed, since λ is bounded by assumption) the cost of the rotated and original problem only differ by the constant $\lambda\left[\rho_0\right]$, for any evolution of the density ρ_k which satisfies the Markov chain (1). Consequently, we obtain $V_\star^{\mathrm{R}}\left[\rho_0\right] = V_\star\left[\rho_0\right] + \lambda\left[\rho_0\right]$.

We now observe that $\mathcal{L}^{\mathrm{R}}$ and $V_\star^{\mathrm{R}}$ satisfy the assumptions of Theorem 1. Consequently, the rotated problem yields a policy guaranteeing that the closed-loop system satisfies D-asymptotic stability with respect to density $\rho_\star$. Because the optimal policies of the rotated and original problem coincide, this proves the second claim. □

The existence of a bounded functional λ satisfying (43) entails the stability of (24) in the set of probability measures. We will show next that such a storage function exists in the LQR case, and can be explicitly provided, hence giving credence to this concept. In line with Lemma 1, the resulting modified cost functional does not take the linear form (42). It is not trivial, and its derivation is fairly technical.

5 The LQR Case

In this section, we will develop a storage functional $\lambda[\rho]$ satisfying (43) for the LQR case with Gaussian process noise and for the Kullback-Leibler divergence D_{KL}. Most proofs are provided in the Appendix. We consider the dynamics:

$$\boldsymbol{s}_+ = A\boldsymbol{s} + B\boldsymbol{a} + \boldsymbol{w}, \tag{47}$$

where $\boldsymbol{s} \in \mathbb{R}^n$ and $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{w}})$ i.i.d., $\mathbb{E}\left[\boldsymbol{w}\boldsymbol{s}^\top\right] = 0$, $\mathbb{E}\left[\boldsymbol{w}\boldsymbol{a}^\top\right] = 0$ and we consider the stage cost

$$L(\boldsymbol{s}, \boldsymbol{a}) = \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{a} \end{bmatrix}^\top H \begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{a} \end{bmatrix}, \quad H = \begin{bmatrix} T & U^\top \\ U & R \end{bmatrix} \succ 0, \tag{48}$$

For $\rho_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$, the dynamics of the system are given by $\rho_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ where the mean and covariance dynamics read as:

$$\boldsymbol{\mu}_{k+1} = A_{\mathrm{c}}\boldsymbol{\mu}_k, \tag{49a}$$

$$\Sigma_{k+1} = A_{\mathrm{c}}\Sigma_k A_{\mathrm{c}}^\top + \Sigma_{\boldsymbol{w}}, \tag{49b}$$

and where $A_{\mathrm{c}} = A - BK$, and K is the regular LQR matrix gain associated to A, B, H. Furthermore, we have

$$\mathbb{E}_{\boldsymbol{s}\sim\rho_k}\left[L\left(\boldsymbol{s}, \boldsymbol{\pi}\left(\boldsymbol{s}\right)\right)\right] = \mathrm{Tr}\left(W\Sigma_k\right) + \boldsymbol{\mu}_k^\top W \boldsymbol{\mu}_k, \tag{50}$$

$$\begin{aligned} D_{\mathrm{KL}}\left(\rho \,||\, \rho_\star\right) = &\frac{1}{2}\left(\mathrm{Tr}\left(\Sigma_\infty^{-1}\Sigma_k\right) + \boldsymbol{\mu}_k^\top\Sigma_\infty^{-1}\boldsymbol{\mu}_k - n \right. \\ &\left. + \log\det\left(\Sigma_\infty\right) - \log\det\left(\Sigma_k\right)\right), \end{aligned}$$

where n is the dimension of $\boldsymbol{s}$,

$$W = \begin{bmatrix} I \\ -K \end{bmatrix}^\top H \begin{bmatrix} I \\ -K \end{bmatrix}, \tag{51}$$

and $\rho_\star$ has the mean $\boldsymbol{\mu}_\infty = 0$ and its variance is the solution of the Lyapunov equation:

$$\Sigma_\infty = A_{\mathrm{c}}\Sigma_\infty A_{\mathrm{c}}^\top + \Sigma_{\boldsymbol{w}}. \tag{52}$$

Note that (52) has a solution only if A_{c} is stabilizing, which also entails that $\boldsymbol{\mu}_\infty = 0$. We observe that for the MDP to be well posed

$$L_0 = \mathrm{Tr}\left(W\Sigma_\infty\right) \tag{53}$$

must hold. We then observe that $\mathcal{L}$, as defined by (25) reads as

$$\mathcal{L}\left[\rho_k, \boldsymbol{\pi}\right] = \boldsymbol{\mu}_k^\top W \boldsymbol{\mu}_k + \mathrm{Tr}\left(W\left(\Sigma_k - \Sigma_\infty\right)\right). \tag{54}$$

One can observe that (54) does not necessarily satisfy (29a), such that a rotation of the LQR stage cost, as per Section 4, is required in order for the optimal cost $V_\star$ associated to the LQR problem, defined according to (24), to be a Lyapunov functional. Before delivering the storage functional associated to the LQR problem, the following section establishes some basic results confirming the convergence of the LQR problem under D_{KL} independently of the proposed theory.

5.1 Convergence under D_{KL}

The stability of the stochastic LQR in the D_{KL} sense can be established by the theory proposed above, via providing a storage functional λ, and as a result a Lyapunov functional for the problem. This is the approach used by the established dissipativity theory in economic MPC, allowing one to build a Lyapunov function directly from the cost and dynamics of the problem at hand. We will present this approach in Section 5.2. Unlike general optimal control and MPC problems, the LQR problem admits a simple and explicit policy, allowing one to explicitly describe the closed-loop dynamics and discuss their stability directly. In this section, we will adopt this approach first and confirm that $D_{\mathrm{KL}}\left(\rho_k || \rho_\star\right)$ is monotonically decreasing under the LQR closed-loop dynamics. This will require several technical results that will also be needed for building a storage functional. The proofs of Lemma 2, Lemma 3, Proposition 1, Lemma 4, and Theorem 3 are provided in the Appendix. In order to proceed, let us introduce first a useful technical lemma.

Lemma 2 *Consider a full-rank matrix $M \in \mathbb{R}^{n\times n}$ such that its maximum singular value, i.e., $\sigma_{\max}(M)$ is less than 1, and consider a symmetric (possibly indefinite) matrix $\Delta \in \mathbb{R}^{n\times n}$. Consider $\Lambda_{1,\ldots,n}(\cdot)$ the ordered eigenvalues of an $\mathbb{R}^{n\times n}$ matrix $\cdot$, where the indexing denotes that order. Then the following holds:*

$$\Lambda_i\left(M\Delta M^\top\right) = \alpha_i\Lambda_i\left(\Delta\right), \quad i = 1, \ldots, n \tag{55}$$

for a sequence $\alpha_i > 0$, $i = 1, \ldots, n$ with $\alpha_i \le \sigma_{\max}(M)$.

□

This lemma will be instrumental in showing the convergence of the LQR problem in the D_{KL} sense, i.e.:

$$D_{\mathrm{KL}}\left(\rho_{k+1} \,||\, \rho_\star\right) < D_{\mathrm{KL}}\left(\rho_k \,||\, \rho_\star\right). \tag{56}$$

In order to obtain this result, the following lemma will be useful, and follows fairly directly from Lemma 2.

Lemma 3 *Under dynamics* (49b)*, the ordered eigenvalues of $\Sigma_\infty^{-1}\Sigma_k$, i.e., $\Lambda_{1,\ldots,n}\left(\Sigma_\infty^{-1}\Sigma_k\right)$ converge monotonically to 1 without changing sign. More specifically:*

$$\Lambda_i\left(\Sigma_\infty^{-1}\Sigma_{k+1}\right) - 1 = \alpha_i\left(\Lambda_i\left(\Sigma_\infty^{-1}\Sigma_k\right) - 1\right), \tag{57}$$

for a sequence $\alpha_{1,\ldots,n} \in \mathbb{R}_+$ with

$$\alpha_i \le \sigma_{\max}\left(\Sigma_\infty^{-\frac{1}{2}}A_{\mathrm{c}}\Sigma_\infty^{\frac{1}{2}}\right) < 1, \qquad \forall i. \tag{58}$$

□

Using Lemma 3, the monotonic decreasing of a class of dissimilarity measures under the dynamics (49) is established next.

Proposition 1 *Consider any dissimilarity measure $D\left(\rho_k \,||\, \rho_\star\right)$ that can be expressed in the form:*

$$D\left(\rho_k \,||\, \rho_\star\right) = c + \boldsymbol{\mu}_k^\top \Sigma_\infty^{-1} \boldsymbol{\mu}_k + \sum_{i=1}^{n} \zeta\left(\Lambda_i\left(\Sigma_\infty^{-1}\Sigma_k\right)\right), \tag{59}$$

for some function $\zeta : \mathbb{R} \to \mathbb{R}_+$ that is strictly increasing away from 1, and some constant c. Then $D\left(\rho_k \,||\, \rho_\star\right)$ is strictly decreasing under dynamics (49).

□

Proposition 1 establishes that in the stochastic LQR case, the measures underlying the stochasticity of the state space converge in the sense of an entire class of dissimilarity measures (including D_{KL}). This is a direct result not relying on the proposed functional dissipativity theory. However, the mathematical argument establishing Proposition 1 is central in developing a storage function showing the functional dissipativity of LQR.

Proposition 1 applies to several dissimilarity measures including D_{KL} and the Wasserstein metric. For the sake of simplicity, we focus on D_{KL} in the following.

Corollary 2 *Dynamics (49) converge monotonically under D_{KL}, i.e.,*

$$D_{\mathrm{KL}}\left(\rho_{k+1} \,||\, \rho_\star\right) \leq D_{\mathrm{KL}}\left(\rho_k \,||\, \rho_\star\right), \tag{60}$$

and the inequalities are strict for $\rho_k \neq \rho_\star$.

□

PROOF. We observe that

$$\begin{aligned} D_{\mathrm{KL}}\left(\rho_k \,||\, \rho_\star\right) = {} & \frac{1}{2}\boldsymbol{\mu}_k^\top \Sigma_\infty^{-1} \boldsymbol{\mu}_k & (61) \\ & + \frac{1}{2}\sum_{i=1}^{n}\left(\Lambda_i\left(\Sigma_\infty^{-1}\Sigma_k\right) - \log \Lambda_i\left(\Sigma_\infty^{-1}\Sigma_k\right) - 1\right), \end{aligned}$$

and we observe that the scalar function

$$\zeta(x) = x - \log x - 1 \tag{62}$$

is monotonically increasing away from 1. Since D_{KL} differs from the form (59) only by a factor $\frac{1}{2}$, the monotonic decrease is conserved. □

5.2 A Storage Functional for D_{KL}

Section 5.1 shows the stability of the stochastic LQR closed-loop trajectories under D_{KL} directly. This can be done in the LQR case where the closed-loop dynamics are known explicitly. For general optimal control and MPC problems, the optimal policy is typically not explicitly known and the same approach cannot be used. Dissipativity theory then allows one to study the stability of a problem based on the cost and dynamics alone, i.e., without using the policy explicitly. This section follows up on dissipativity theory and shows the existence of a storage functional for the LQR case, hence illustrating the theory presented in this paper. We need to start with the following technical lemma.

Lemma 4 *Consider the function:*

$$\varsigma\left(\Delta\right) = \mathrm{Tr}\left(\Delta\right) - \log\det\left(\Delta + I\right). \tag{63}$$

For any symmetric matrix Δ and matrix M such that $\sigma_{\max}\left(M\right) < 1$, the following inequality holds:

$$\left(1 - \beta\right)\varsigma\left(\Delta\right) - \varsigma\left(M\Delta M^\top\right) \geq 0, \tag{64}$$

for any $\beta \leq 1 - \sigma_{\max}\left(M\right)$.

□

Equipped with this lemma, we are now ready to provide a storage functional for the LQR case.

Theorem 3 *The choice:*

$$\begin{aligned} \lambda\left[\rho_k\right] = {} & \kappa\left(\mathrm{Tr}\left(\Sigma_\infty^{-1}\Sigma_k\right) + \log\det\left(\Sigma_\infty^{-1}\Sigma_k\right) - n\right) \\ & - \mathrm{Tr}\left(\Omega\left(\Sigma_\infty^{-\frac{1}{2}}\Sigma_k\Sigma_\infty^{-\frac{1}{2}} - I\right)\right) \end{aligned} \tag{65}$$

satisfies the functional dissipativity criterion (43) where matrix Ω is solution of the discrete Lyapunov equation:

$$M\Omega M^\top - \Omega + \Sigma_\infty^{\frac{1}{2}} W \Sigma_\infty^{\frac{1}{2}} = 0, \tag{66}$$

for $M = \Sigma_\infty^{-\frac{1}{2}} A_{\mathrm{c}} \Sigma_\infty^{\frac{1}{2}}$ and where κ, ϱ are constants satisfying:

$$\kappa \geq \frac{1}{2\left(1 - \sigma_{\max}\left(M\right)\right)}, \quad \varrho \leq 2\sigma_{\min}\left(W\right)\sigma_{\min}\left(\Sigma_\infty\right). \tag{67}$$

□

Remark 7 *It is useful to note here that (65) cannot be cast as a linear functional of ρ_k. This precludes forms like (10), and confirms the arguments made in the first part of this paper.*

5.3 Illustration

We illustrate next Theorem 1-3 for the LQR case. We chose a case with $n = 2$ states, i.e., $\boldsymbol{s}_k \in \mathbb{R}^2$, and a scalar action $\boldsymbol{a}_k \in \mathbb{R}$ having the dynamics

$$\boldsymbol{s}_{k+1} = \frac{1}{10}\begin{bmatrix} 8 & 5 \\ -5 & 7 \end{bmatrix}\boldsymbol{s}_k + \frac{1}{10}\begin{bmatrix} 0 \\ 5 \end{bmatrix}\boldsymbol{a}_k + \boldsymbol{w}_k, \tag{68}$$

where $\boldsymbol{w}_k \sim \mathcal{N}\left(0, \Sigma_{\boldsymbol{w}}\right)$ and

$$\Sigma_{\boldsymbol{w}} = \begin{bmatrix} 2 & -1 \\ -1 & 1.6 \end{bmatrix}. \tag{69}$$

The stage cost is based on the weighting matrices $T = I, R = 1, U = 0$ in (48). The corresponding constant L_0 solution of (5) reads as:

$$L_0 = 8.92. \tag{70}$$

The initial density $\rho_0 = \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0\right)$ was selected, with

$$\boldsymbol{\mu}_0 = 1.6\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \Sigma_0 = \frac{1}{10}\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}. \tag{71}$$

The optimal steady-state density $\rho_\star = \mathcal{N}\left(\mathbf{0}, \Sigma_\infty\right)$ has the covariance

$$\Sigma_\infty = \begin{bmatrix} 3.73 & -1.76 \\ -1.76 & 3 \end{bmatrix}. \tag{72}$$

The constants $\kappa = 1.72$, $\varrho = 3.14$ then satisfy (67). Figure 1 provides a graphical illustration of this case, showing the trajectories $\rho_k = \mathcal{N}\left(\boldsymbol{\mu}_k, \Sigma_k\right)$, in terms of their mean (center of the ellipsoids in red) and covariance (ellipsoids). For $k \to \infty$, the densities converge to the steady-state optimal density $\rho_\star$, depicted as the green ellipsoid here, starting from the initial density ρ_0 represented as the light black ellipsoid.

Figure 2 upper graph shows the stage cost $\mathcal{L}\left[\rho_k, \boldsymbol{\pi}\right]$ given by (54). We observe that $\mathcal{L}$ does not satisfy (29a) as it can take negative values, hence it cannot be used to establish stability as it violates the assumptions of Theorem 1. The lower graph depicts D_{KL} (scaled by a factor φ) and the rotated cost $\mathcal{L}^{\mathrm{R}}\left[\rho_k, \boldsymbol{\pi}\right]$ (41), using the storage functional (65) prescribed by Theorem 3. One can observe how $\mathcal{L}^{\mathrm{R}}$ selected as per Theorem 3 is lower-bounded by the scaled D_{KL} and satisfies (43). As a result, it satisfies the conditions of stability (29a) of Theorem 1. One can also observe how D_{KL} satisfies Proposition 1.

Figure 3 illustrates the evolution of the value function $J_\star$ from the original MDP (4) (see red curve). One can readily see that $J_\star$ is not a Lyapunov function as it is not decreasing. The cyan curve represents the evolution of the value function $V_\star^{\mathrm{R}}$ of the rotated problem (44), which decreases monotonically, as established by Theorem 2.

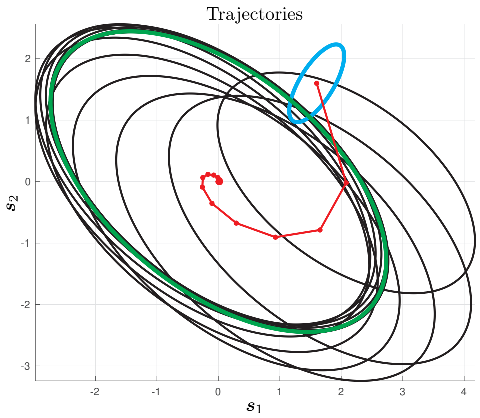

Fig. 1. Illustration of the LQR case. The expected closed-loop trajectories are depicted in red and the 1σ ellipsoids in black. The optimal steady-state density $\rho_\star$ is depicted as the green ellipsoid, and the initial density ρ_0 as the black ellipsoid. One can observe how the system converges to $\rho_\star$.

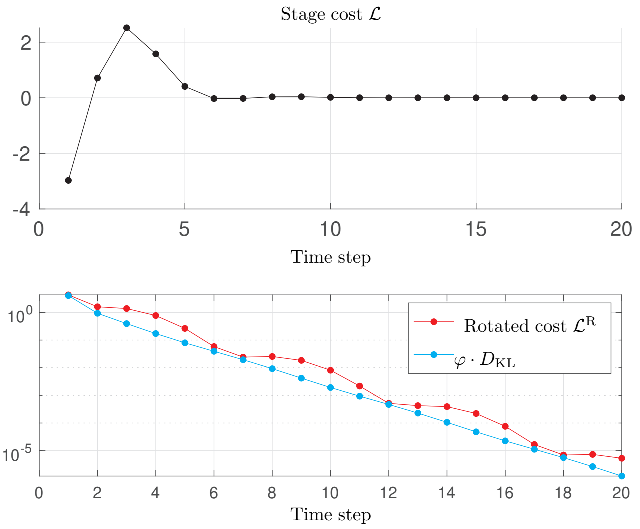

Fig. 2. Illustration of the LQR case. The upper graph depicts the stage cost (54). The lower graph depicts $\varphi \cdot D_{\mathrm{KL}}$ and the rotated cost $\mathcal{L}^{\mathrm{R}}$ (41), using the storage functional (65), using $\varphi = 3.13$, selected according to (67).

Figure 4 illustrates assumption (45) in Theorem 2, and similarly assumption (29b) in Theorem 1. As detailed above, these assumptions are difficult to verify formally even in the deterministic case. One can observe, however, that they hold on this specific example and on the proposed trajectories.

6 Conclusion

In this paper, we proposed a generalization of the established economic MPC dissipativity theory to Markov Decision Processes. We explain why this generalization is not straightforward, and show that it can be done by extending the notion of storage functions to nonlinear storage functionals. A classic Lyapunov argument can

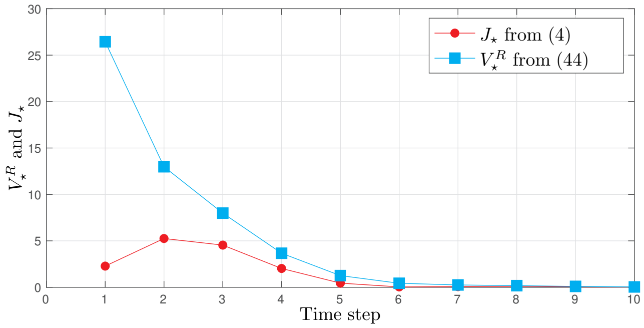

Fig. 3. Illustration of the LQR case. The red curve represents the evolution of the value function $J_\star$ from the original problem (4). The black curve represents the evolution of the value function $V_\star^{\mathrm{R}}$ of the rotated problem (44), which is a Lyapunov function for the system, as established in Theorem 2.

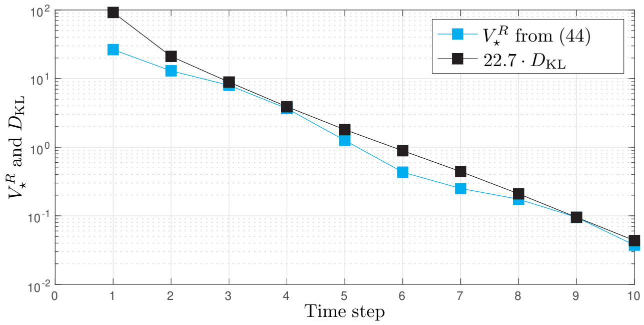

Fig. 4. Illustration of the LQR case. The black curve represents the evolution of the value function $V_\star^{\mathrm{R}}$ of the rotated problem (44). The black curve is D_{KL} scaled by a factor 22.7, and is upper-bounding $V_\star^{\mathrm{R}}$, illustrating assumption (45) in Theorem 2 on these specific trajectories and specific case.

then be used to discuss the asymptotic stability of the probability measures underlying the Markov Decision Processes to the steady-state optimal probability measure. The asymptotic stability can be expressed in terms of dissimilarity measures such as the Kullback-Leibler divergence, the Wasserstein metric, or the total variational distance. The theory is illustrated on the LQR case with Normal process noise, for which a storage functional can be explicitly provided.

Current work extends this theory to discussing the stability of Stochastic MPC, the construction of stability-constrained learning using MPC, and the extension to the discounted case, by building on the theory proposed in [22]. Furthermore, an extension to policies based on finite-horizon MPC schemes will be considered.

References

[1] R. Amrit, J. Rawlings, and D. Angeli. Economic Optimization Using Model Predictive Control with a Terminal Cost. *Annual Reviews in Control*, 35:178–186, 2011.

[2] Florian A. Bayer, Matthias Lorenzen, Matthias A. Müller, and Frank Allgöwer. Robust Economic Model Predictive Control Using Stochastic Information. *Automatica*, 74:151–161, 2016.

[3] Florian A. Bayer, Matthias A. Müller, and Frank Allgöwer. On optimal system operation in robust economic mpc. *Automatica*, 88:98–106, 2018.

[4] L. Chisci, J.A. Rossiter, and G. Zappa. Systems with persistent disturbances: predictive control with restricted constraints. *Automatica*, 37:1019–1028, 2001.

[5] M. Diehl, R. Amrit, and J.B. Rawlings. A Lyapunov Function for Economic Optimizing Model Predictive Control. *IEEE Transactions of Automatic Control*, 56(3):703–707, March 2011.

[6] T. Faulwasser, L. Grüne, and M. Müller. Economic nonlinear model predictive control: Stability, optimality and performance. *Foundations and Trends in Systems and Control*, 5(1):1–98, 2018.

[7] S. Gros. An Economic NMPC Formulation for Wind Turbine Control. In *Conference on Decision and Control*, 2013.

[8] L. Grüne. Economic receding horizon control without terminal constraints. *Automatica*, 49:725–734, 2013.

[9] L. Grüne and J. Pannek. *Nonlinear Model Predictive Control*. Communications and Control Engineering. Springer International Publishing, 2 edition, 2017.

[10] R. Hult, M. Zanon, S. Gros, and P. Falcone. Energy-Optimal Coordination of Autonomous Vehicles at Intersections. In *2018 European Control Conference (ECC)*, pages 602–607, June 2018.

[11] Hans W. Knobloch. Disturbance Attenuation in Control Systems. *International Game Theory Review*, 07(03):261–283, 2005.

[12] D.Q. Mayne, M.M. Seron, and S.V. Rakovic. Robust Model Predictive Control of Constrained Linear Systems with Bounded Disturbances. *Automatica*, 41:219–224, 2005.

[13] Ali Mesbah. Stochastic Model Predictive Control: An Overview and Perspectives for Future Research. *IEEE Control Systems Magazine*, 36(6):30–44, 2016.

[14] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, USA, 2nd edition, 2009.

[15] M. A. Müller, D. Angeli, and F. Allgöwer. On Necessity and Robustness of Dissipativity in Economic Model Predictive Control. *IEEE Transactions on Automatic Control*, 60(6):1671–1676, 2015.

[16] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.

[17] Tanmay Rajpurohit and Wassim M. Haddad. Dissipativity Theory for Nonlinear Stochastic Dynamical Systems. *IEEE Transactions on Automatic Control*, 62(4):1684–1699, 2017.

[18] James B. Rawlings and Rishi Amrit. Optimizing Process Economic Performance using Model Predictive Control. In *Proceedings of NMPC 08 Pavia*, pages 119–138. 2009.

[19] James B. Rawlings, David Q. Mayne, and Moritz Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2 edition, 2017.

[20] Pantelis Sopasakis, Domagoj Herceg, Panagiotis Patrinos, and Alberto Bemporad. Stochastic Economic Model Predictive Control for Markovian Switching Systems. *IFAC-PapersOnLine*, 50(1):524–530, 2017. 20th IFAC World Congress.

[21] J. Vasilj, S. Gros, D. Jakus, and M. Zanon. Day-Ahead Scheduling and Real-Time Economic MPC of CHP Unit in

Microgrid With Smart Buildings. *IEEE Transactions on Smart Grid*, 10(2):1992–2001, March 2019.

[22] M. Zanon and S. Gros. A New Dissipativity Condition for Asymptotic Stability of Discounted Economic MPC. *Automatica.* (submitted) https://arxiv.org/abs/2106.09377.

[23] M. Zanon and S. Gros. On the Similarity Between Two Popular Tube MPC Formulations. In *Proceedings of the European Control Conference*, 2021. (accepted).

[24] M. Zanon, L. Grüne, and M. Diehl. Periodic Optimal Control, Dissipativity and MPC. *IEEE Transactions on Automatic Control*, 62(6):2943–2949, 2017.

7 Appendix

The proofs of Lemma 2, Lemma 3, Proposition 1, Lemma 4, and Theorem 3 are provided hereafter. Note that we will denote the ordered eigenvalues of matrix A as $\Lambda_{1,\dots,n}(A)$ and its ordered singular values as $\sigma_{1,\dots,n}(A)$. We will denote the trace of A as $\mathrm{Tr}\,(A)$. We will further use:

$$\begin{aligned}
\Lambda_i(AB) &= \Lambda_i(BA), & 1 + c\Lambda_i(A) &= \Lambda_i(I + cA),\\
\Lambda_i(AA^\top) &= \Lambda_i(A)^2, & \max_k \sigma_k(A)\sigma_i(B) &\geq \sigma_i(AB),\\
\mathrm{Tr}(ABC) &= \mathrm{Tr}(CAB), & \det(A)\det(B) &= \det(AB),\\
\det(ABC) &= \det(CAB)
\end{aligned}$$

and all such permutations

7.1 Proof of Lemma 2

First, we observe that the following inequality holds:

$$\sigma\left(M\Delta M^\top\right) \leq \sigma_{\max}\left(M\right)^2 \sigma\left(\Delta\right), \tag{73}$$

where $\sigma\left(\cdot\right)$ is the vector of singular values. Hence since Δ is symmetric:

$$\Lambda_i\left(M\Delta M^\top\right)^2 \leq \sigma_{\max}\left(M\right)^2 \Lambda_i\left(\Delta\right)^2, \tag{74}$$

such that

$$\left|\Lambda_i\left(M\Delta M^\top\right)\right| = |\alpha_i|\left|\Lambda_i\left(\Delta\right)\right| \tag{75}$$

holds for some sequence $|\alpha_1|, \dots, |\alpha_n| \leq \sigma_{\max}\left(M\right)$. We then need to show that $\alpha_i > 0$, $i = 1, \dots, n$. We observe that:

$$\Lambda_i\left(M\Delta M^\top\right) = \Lambda_i\left(\Phi\Delta\right), \tag{76}$$

where $\Phi = M^\top M$ is symmetric, positive definite. Let us define:

$$\Gamma(t) = e^{t\log\Phi}\Delta, \tag{77}$$

where we use the matrix exponential and logarithms. Then

$$\Gamma(0) = \Delta \quad \text{and} \quad \Gamma(1) = M^\top M\Delta \tag{78}$$

trivially hold. We further observe that

$$\det\left(\Gamma(t)\right) = \det\left(e^{t\log\Phi}\right)\det\left(\Delta\right) = e^{t\mathrm{Tr}(\log\Phi)}\det\left(\Delta\right) \tag{79}$$

$$= \left(e^{\mathrm{Tr}(\log\Phi)}\right)^t \det\left(\Delta\right) = \det\left(\log\Phi\right)^t \det\left(\Delta\right).$$

Since $\det\left(\Gamma(1)\right) = \det\left(M\Delta M^\top\right) \neq 0$ by assumption, it follows that $\det\left(\Gamma(t)\right) \neq 0$ for all $t \in [0, 1]$. We can then conclude that the eigenvalues

$$\Lambda_i\left(\Gamma(t)\right) \neq 0, \quad \forall\, t, \tag{80}$$

such that $\Lambda_i\left(\Gamma(t)\right)$ does change sign over $t \in [0, 1]$. This establishes that $\alpha_i > 0$, $i = 1, \dots, n$ and hence (55).

7.2 Proof of Lemma 3

We first observe that the dynamics for Σ_k can be reformulated as:

$$\Psi_{k+1} = M\Psi_k M^\top + N, \tag{81}$$

where $\Psi_k = \Sigma_\infty^{-\frac{1}{2}}\Sigma_k\Sigma_\infty^{-\frac{1}{2}}$ and

$$M = \Sigma_\infty^{-\frac{1}{2}} A_{\mathrm{c}} \Sigma_\infty^{\frac{1}{2}}, \qquad N = \Sigma_\infty^{-\frac{1}{2}}\Sigma_{\boldsymbol{w}}\Sigma_\infty^{-\frac{1}{2}}, \tag{82}$$

such that $\lim_{k\to\infty}\Psi_k = I$ and

$$MM^\top + N = I. \tag{83}$$

We can then observe that

$$\begin{aligned}
\sigma_i\left(M\right)^2 = \Lambda_i\left(M^\top M\right) = \Lambda_i\left(MM^\top\right) = \Lambda_i\left(I - N\right)\\
= 1 - \Lambda_i(N) \geq 0,
\end{aligned} \tag{84}$$

and that

$$\Lambda_i\left(N\right) = \Lambda_i\left(\Sigma_\infty^{-\frac{1}{2}}\Sigma_{\boldsymbol{w}}\Sigma_\infty^{-\frac{1}{2}}\right) = \Lambda_i\left(\Sigma_\infty^{-1}\Sigma_{\boldsymbol{w}}\right) > 0, \tag{85}$$

since Σ_∞, $\Sigma_{\boldsymbol{w}}$ are positive definite. It follows that

$$\sigma_i\left(M\right) \in [0, 1). \tag{86}$$

Let us label

$$\Delta_k = \Psi_k - I, \tag{87}$$

and observe that

$$\Delta_{k+1} = M\Delta_k M^\top \quad \text{and} \quad \Lambda_i\left(\Delta_k\right) = \Lambda\left(\Psi_k\right) - 1. \tag{88}$$

Using Lemma 2, we observe that (55) applies, i.e.

$$\Lambda_i\left(\Delta_{k+1}\right) = \alpha_i \Lambda_i\left(\Delta_k\right), \quad i = 1, \ldots, n, \tag{89}$$

for a sequence $\alpha_i > 0$, $i = 1, \ldots, n$ with $\alpha_i \leq \sigma_{\max}\left(M\right) < 1$. Finally, we observe that:

$$\Lambda_i\left(\Delta_k\right) = \Lambda_i\left(\Sigma_\infty^{-\frac{1}{2}} \Sigma_k \Sigma_\infty^{-\frac{1}{2}} - I\right) = \Lambda_i\left(\Sigma_\infty^{-1}\Sigma_k\right) - 1, \tag{90}$$

and conclude that (57) holds.

7.3 Proof of Proposition 1

We first observe that the monotonic convergence of the second term in (59)

$$\sum_{i=1}^{n} \zeta\left(\Lambda_i\left(\Sigma_\infty^{-1}\Sigma_k\right)\right) \tag{91}$$

follows directly from Lemma 3. The convergence of the first term follows classic system dynamic theory. We recall the argument for completeness. Consider the state space transformation:

$$\boldsymbol{\nu}_k = \Sigma_\infty^{-\frac{1}{2}} \boldsymbol{\mu}_k, \tag{92}$$

following the dynamics:

$$\boldsymbol{\nu}_{k+1} = M \boldsymbol{\nu}_k, \tag{93}$$

with

$$\Lambda_i\left(M^\top M\right) \leq \sigma_{\max}\left(M\right)^2 < 1. \tag{94}$$

It follows that

$$\begin{aligned} \boldsymbol{\mu}_{k+1}^\top \Sigma_\infty^{-1} \boldsymbol{\mu}_{k+1} = \|\boldsymbol{\nu}_{k+1}\|^2 = \boldsymbol{\nu}_k^\top M^\top M \boldsymbol{\nu}_k < \|\boldsymbol{\nu}_k\|^2 \\ = \boldsymbol{\mu}_k^\top \Sigma_\infty^{-1} \boldsymbol{\mu}_k. \end{aligned} \tag{95}$$

7.4 Proof of Lemma 4

We first observe that for any $a < 1$ the inequality:

$$\begin{aligned} &\vartheta_{a,b}(x) \\ &:= \left(1 - b\right)\left(x - \log\left(x + 1\right)\right) - ax + \log\left(ax + 1\right) \geq 0 \end{aligned} \tag{96}$$

holds on $x \in (-1, \infty)$ for $0 < b \leq 1 - a < 1$. Indeed, we observe that $\vartheta_{a,b}(0) = 0$ and that on the interval $x \in (-1, \infty)$

$$\frac{\mathrm{d}\vartheta_{a,b}}{\mathrm{d}x} = -\frac{ax\left(\left(a + b - 1\right)x - 1 + b\right)}{\left(ax + 1\right)\left(x + 1\right)} = 0 \tag{97}$$

has the unique solution $x = 0$. Furthermore, the sign of $\frac{\mathrm{d}\vartheta_{a,b}}{\mathrm{d}x}$ entails that $\vartheta_{a,b}$ is monotonically increasing away from $x = 0$, which establishes (96). Using Lemma 2, we then observe that for all i:

$$\Lambda_i\left(M\Delta M^\top\right) = \Lambda_i\left(M^\top M\Delta\right) = \alpha_i \Lambda_i\left(\Delta\right) \tag{98}$$

holds for some sequence $\alpha_{1,\ldots,n} \leq \sigma_{\max}\left(M\right)$. Then

$$\begin{aligned} &-\Lambda_i\left(M\Delta M^\top\right) + \log\left(\Lambda_i\left(M\Delta M^\top\right) + 1\right) = \\ &-\alpha_i\Lambda_i\left(\Delta\right) + \log\left(\alpha_i\Lambda_i\left(\Delta\right) + 1\right). \end{aligned} \tag{99}$$

Hence, using

$$\begin{aligned} \varsigma\left(\Delta\right) &= \sum_{i=1}^{n} \Lambda_i\left(\Delta\right) - \log\left(\Lambda_i\left(\Delta\right) + 1\right) \\ \varsigma\left(M\Delta M^\top\right) &= \sum_{i=1}^{n} \Lambda_i\left(M\Delta M^\top\right) \\ &\qquad - \log\left(\Lambda_i\left(M\Delta M^\top\right) + 1\right), \end{aligned} \tag{100}$$

we observe that

$$\left(1 - \beta\right)\varsigma\left(\Delta\right) - \varsigma\left(M\Delta M^\top\right) = \sum_{i=1}^{n} \vartheta_{\alpha_i,\beta}(\Lambda_i\left(\Delta\right)),$$

such that the choice

$$\beta \leq \min_i 1 - \alpha_i \leq 1 - \sigma_{\max}\left(M\right) < 1 \tag{101}$$

ensures that (64) holds.

7.5 Proof of Theorem 3

We first observe that $\lambda\left[\rho_\star\right] = 0$ holds by construction. We further observe that:

$$\mathrm{Tr}\left(\Omega\left(\Sigma_\infty^{-\frac{1}{2}}\Sigma_k\Sigma_\infty^{-\frac{1}{2}} - I\right)\right) = \mathrm{Tr}\left(\Omega\Delta_k\right), \tag{102}$$

and, using (88) and (66), we obtain:

$$\mathrm{Tr}\left(\Omega\Delta_k\right) - \mathrm{Tr}\left(\Omega\Delta_{k+1}\right) = -\mathrm{Tr}\left(W\left(\Sigma_k - \Sigma_\infty\right)\right). \tag{103}$$

Using (103) and (63) we then observe that:

$$\begin{aligned} \lambda\left[\rho_k\right] - \lambda\left[\rho_{k+1}\right] = \kappa(\varsigma\left(\Delta_k\right) - \varsigma\left(\Delta_{k+1}\right)) \\ - \mathrm{Tr}\left(W\left(\Sigma_k - \Sigma_\infty\right)\right). \end{aligned} \tag{104}$$

Using (54) and (104) and Lemma 4, it follows that

$$\begin{aligned} &\mathcal{L}\left[\rho_k, \boldsymbol{\pi}[\rho_k]\right] - \lambda\left[\rho_{k+1}\right] + \lambda\left[\rho_k\right] \\ &\quad = \boldsymbol{\mu}_k^\top W \boldsymbol{\mu}_k + \kappa(\varsigma\left(\Delta_k\right) - \varsigma\left(\Delta_{k+1}\right)) \\ &\quad \geq \boldsymbol{\mu}_k^\top W \boldsymbol{\mu}_k + \kappa\beta\varsigma\left(\Delta_k\right). \end{aligned} \tag{105}$$

We further observe that

$$D_{\mathrm{KL}}\left(\rho_k \,||\, \rho_\star\right) = \frac{1}{2}\boldsymbol{\mu}_k^\top \Sigma_\infty^{-1} \boldsymbol{\mu}_k + \frac{1}{2}\varsigma\left(\Delta_k\right). \tag{106}$$

It follows that for (67), $\kappa \geq \frac{1}{2\beta}$, and the inequality

$$\mathcal{L}\left[\rho_k, \boldsymbol{\pi}[\rho_k]\right] - \lambda\left[\rho_{k+1}\right] + \lambda\left[\rho_k\right] \geq \varrho D_{\mathrm{KL}}\left(\rho_k \,||\, \rho_\star\right) \tag{107}$$

holds.