

Equivalence of Optimality Criteria for Markov Decision Process and Model Predictive Control

Arash Bahari Kordabad, Mario Zanon, Sebastien Gros

Abstract—This paper shows that the optimal policy and value functions of a Markov Decision Process (MDP), either discounted or not, can be captured by a finite-horizon undiscounted Optimal Control Problem (OCP), even if based on an inexact model. This can be achieved by selecting a proper stage cost and terminal cost for the OCP. A very useful particular case of OCP is a Model Predictive Control (MPC) scheme where a deterministic (possibly nonlinear) model is used to reduce the computational complexity. This observation leads us to parameterize an MPC scheme fully, including the cost function. In practice, Reinforcement Learning algorithms can then be used to tune the parameterized MPC scheme. We verify the developed theorems analytically in an LQR case and we investigate some other nonlinear examples in simulations.

Index Terms—Markov Decision Process, Model Predictive Control, Reinforcement Learning, Optimality

I. INTRODUCTION

Markov Decision Processes (MDPs) provide a standard framework for the optimal control of discrete-time stochastic processes, where the stage cost and transition probability depend only on the current state and the current input of the system [1]. A control system, described by an MDP, receives an input at each time instance and proceeds to a new state with a given probability density, and in the meantime, it gets a stage cost at each transition. For an MDP, a policy is a mapping from the state space into the input space and determines how to select the input based on the observation of the current state. This policy can either be a deterministic mapping from the state space or a conditional probability of the current state, describing the stochastic policy [2]. This paper focuses on deterministic policies. Solving an MDP refers to finding an optimal policy that minimizes the expected value of a total cumulative cost as a function of the current state. The cumulative cost can be either discounted or undiscounted with respect to the time instant. Therefore, different definitions for the cumulative cost yields different optimality criteria for the MDPs. Dynamic Programming (DP) techniques can be used to solve MDPs based on the Bellman equations. However, solving the Bellman equations is typically intractable unless the problem is of very low dimension [3]. This issue is known as “curse of dimensionality” in the literature [4]. Besides, DP requires the exact transition probability of MDPs, while in most engineering applications, we do not have access to the exact probability transition of the real system.

Reinforcement Learning (RL) [5] and approximate DP [6] are two common techniques that tackle these difficulties. RL offers powerful tools for tackling MDP without having an accurate knowledge of the probability distribution underlying the state transition. In most cases, RL requires a function approximator to capture the optimal policy or the optimal value functions underlying the MDP. A common choice of function approximator in the RL community is to use a Deep Neural Network (DNN) [7]. DNNs can be used to capture either the optimal policy underlying the MDP directly or the action-value

function from which the optimal policy can be indirectly extracted. However, the formal analysis of closed-loop stability and safety of the policies provided by approximators such as DNNs is challenging. Moreover, DNNs usually need a large number of tunable parameters and a pre-training is often required so that the initial values of the parameters are reasonable.

Model Predictive Control (MPC) is a well-known control strategy that employs a (possibly inaccurate) model of the real system dynamics to produce an input-state sequence over a given finite-horizon such that the resulting predicted state trajectory minimizes a given cost function while explicitly enforcing the input-state constraints imposed on the system trajectories [8]. For computational reasons, simple models are usually preferred in the MPC scheme. Hence, the MPC model often does not have the structure required to correctly capture the real system dynamics and stochasticity. The idea of using MPC as a function approximator for RL techniques was justified first in [9], where it was shown that the optimal policy of a discounted MDP can be captured by a discounted MPC scheme even if the model is inexact. Recently, MPC has been used in different systems to deliver a structured function approximator for MDPs (see e.g., [9]–[11]) and partially observable MDPs [12]. Stability for discounted MPC schemes is challenging, and for a finite-horizon problem, it is shown in [13] that even if the provided stage cost, terminal cost and terminal set satisfy the stability requirements, the closed-loop might be unstable for some discount factors. Indeed, the discount factor has a critical role in the stability of the closed-loop system under the optimal policy of the discounted cost. The conditions for the asymptotic stability for discounted optimal control problems have been recently developed in [14] for deterministic systems with the exact model. Therefore, an undiscounted MPC scheme is more desirable, where the closed-loop stability analysis is straightforward and well-developed [8].

The equivalence of MDPs criteria (discounted and undiscounted) has been recently discussed in [15] in the case an exact model of MDP is available. However, in practice, the exact probability transition of the MDP might not be available and we usually have a (possibly inaccurate) model of the real system. This work extends the results of [15] in the sense of the model mismatch and while extends also the results of [9] to the case of using undiscounted MPC scheme to capture a (possibly discounted) MDP. More specifically, we show that, under some conditions, an undiscounted finite-horizon Optimal Control Problem (OCP) can capture the optimal policy and the optimal value functions of a given MDP, either discounted or undiscounted, even if an inexact model is used in the undiscounted OCP. We then propose to use a deterministic (possibly nonlinear) MPC scheme as a particular case of the theorem to formulate the undiscounted OCP as a common MPC scheme. By parameterizing the MPC scheme, and tuning the parameters via RL algorithms one can achieve the best approximation of the optimal policy and the optimal value functions of the original MDP within the adopted MPC structure.

The paper is structured as follows. Section II provides the formulation of MDPs under discounted and undiscounted optimality criteria. Section III provides formal statements showing that using cost modification in a finite-horizon undiscounted OCP one is able to capture the optimal value function and optimal policy function of the real system with discounted and undiscounted cost even with a wrong

model. Section IV presents a parameterized MPC scheme as a special case of the undiscounted OCP, where the model is deterministic (i.e. the probability transition is a Dirac measure). Then the parameters can be tuned using RL techniques. Section V provides an analytical LQR example. Section VI illustrates different numerical simulation. Finally, section VII delivers the conclusions.

II. REAL SYSTEM

In this section, we formulate the real system as Markov Decision Processes (MDPs). We consider an MDP on a continuous state and input spaces over \mathbb{R}^n and \mathbb{R}^m , respectively, with stochastic states $\mathbf{s}_k \in \mathcal{X} \subseteq \mathbb{R}^n$ in the Lebesgue-measurable set \mathcal{X} and inputs $\mathbf{a}_k \in \mathcal{U} \subseteq \mathbb{R}^m$. The triple $(\mathcal{X}, \mathcal{F}, \rho)$ defines the probability space associated with a Markov chain with associated σ -field \mathcal{F} and ρ is the probability measure. We then consider stochastic dynamics defined by the following conditional probability measure:

$$\rho[\mathbf{s}_{k+1} | \mathbf{s}_k, \mathbf{a}_k], \quad (1)$$

defining the conditional probability of observing a transition from a given state-action pair $\mathbf{s}_k, \mathbf{a}_k$ to a subsequent state \mathbf{s}_{k+1} . The input \mathbf{a} applied to the system for a given state \mathbf{s} is selected by a deterministic policy $\pi : \mathcal{X} \rightarrow \mathcal{U}$. We denote $\mathbf{s}_{0,1,\dots}^\pi$ the (possibly stochastic) trajectories of the system (1) under policy π , i.e., $\mathbf{s}_{k+1}^\pi \sim \rho[\cdot | \mathbf{s}_k^\pi, \pi(\mathbf{s}_k^\pi)]$, starting from $\mathbf{s}_0^\pi = \mathbf{s}$. We further denote the measure associated with such trajectories as τ_k^π in the same space as ρ . More specifically, $\tau_0^\pi(\cdot) = \rho_0(\cdot), \forall \pi$, where $\rho_0(\cdot)$ is the initial state distribution and $\tau_{k+1}^\pi(\cdot) := \int_{\mathcal{X}} \rho[\cdot | \mathbf{s}, \pi(\mathbf{s})] \tau_k^\pi(d\mathbf{s}), k > 0$.

A. Discounted MDPs

In the discounted setting, we aim to find the optimal policy π^* , solution of the following discounted infinite-horizon OCP:

$$V^*(\mathbf{s}) := \min_{\pi} V^\pi(\mathbf{s}) := \mathbb{E}_{\tau^\pi} \left[\sum_{k=0}^{\infty} \gamma^k \ell(\mathbf{s}_k^\pi, \pi(\mathbf{s}_k^\pi)) \right], \quad (2)$$

for all initial states $\mathbf{s}_0^\pi = \mathbf{s}$, where $V^* : \mathcal{X} \rightarrow \mathbb{R}$ is the optimal value function, V^π is the value function of the Markov Chain in closed-loop with policy π , $\ell : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ is the stage cost function of the real system and $\gamma \in (0, 1]$ is the discount factor. The expectation \mathbb{E}_{τ^π} is taken over the distribution underlying the Markov Chain (1) in closed-loop with policy π , i.e., $\mathbf{s}_k \sim \tau_k^\pi(\cdot)$ for $k > 0$. The action-value function $Q^*(\mathbf{s}, \mathbf{a})$ and advantage function $A^*(\mathbf{s}, \mathbf{a})$ associated to (2) are defined as follows:

$$Q^*(\mathbf{s}, \mathbf{a}) := \ell(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\rho} \left[V^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a} \right], \quad (3a)$$

$$A^*(\mathbf{s}, \mathbf{a}) := Q^*(\mathbf{s}, \mathbf{a}) - V^*(\mathbf{s}). \quad (3b)$$

Then from the Bellman equation, we have the following identities:

$$V^*(\mathbf{s}) = Q^*(\mathbf{s}, \pi^*(\mathbf{s})) = \min_{\mathbf{a}} Q^*(\mathbf{s}, \mathbf{a}), \quad \forall \mathbf{s} \in \mathcal{X}, \quad (4a)$$

$$0 = \min_{\mathbf{a}} A^*(\mathbf{s}, \mathbf{a}), \quad \pi^*(\mathbf{s}) \in \arg \min_{\mathbf{a}} A^*(\mathbf{s}, \mathbf{a}), \quad \forall \mathbf{s} \in \mathcal{X}. \quad (4b)$$

B. Undiscounted MDPs

Undiscounted MDPs refer to MDPs when $\gamma = 1$. In this case V^* is in general unbounded and the MDP is ill-posed. In order to tackle this issue, alternative optimality criteria are needed. Gain optimality is one of the common criteria in the undiscounted setting. Gain optimality is defined based on the following average-cost problem:

$$\bar{V}^*(\mathbf{s}) := \min_{\pi} \lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\tau^\pi} \left[\sum_{k=0}^{N-1} \ell(\mathbf{s}_k^\pi, \pi(\mathbf{s}_k^\pi)) \right], \quad (5)$$

for all initial states $\mathbf{s}_0^\pi = \mathbf{s}, \forall \pi$, where \bar{V}^* is the optimal average cost. We denote the optimal policy solution of (5) as $\bar{\pi}^*$. This optimal policy is called *gain optimal*. The gain optimal policy $\bar{\pi}^*$ may not be unique. Moreover, the optimal average cost \bar{V}^* is commonly assumed to be independent of the initial state \mathbf{s} [16]. This assumption e.g. holds for *unichain* MDPs, in which under any policy any state can be reached in finite time from any other state. Unfortunately, the gain optimality criterion only considers the optimal steady-state distribution and it overlooks transients. As an alternative, *bias optimality* considers the optimality of the transients. Precisely, bias optimality can be formulated through the following OCP:

$$\tilde{V}^*(\mathbf{s}) = \min_{\pi} \mathbb{E}_{\tau^\pi} \left[\sum_{k=0}^{\infty} (\ell(\mathbf{s}_k^\pi, \pi(\mathbf{s}_k^\pi)) - \bar{V}^*) \right], \quad (6)$$

where \tilde{V}^* is the optimal value function associated to bias optimality. Note that (6) can be seen as a special case of the discounted setting in (2) when $\gamma = 1$ and the optimal average cost \bar{V}^* is subtracted from the stage cost in (2). Therefore, for the rest of the paper we will consider the discounted setting (2). Without loss of generality we assume that $\bar{V}^* = 0$ in the case $\gamma = 1$. This choice yields a well-posed optimal value function in the undiscounted setting. Clearly, if this does not hold, one can shift the stage cost to achieve $\bar{V}^* = 0$.

III. MODEL OF THE SYSTEM

In general, we may not have full knowledge of the probability transition of the real MDP (1). One then typically considers an imperfect model of the real MDP (1), having the state transition:

$$\hat{\rho}[\mathbf{s}_{k+1} | \mathbf{s}_k, \mathbf{a}_k], \quad (7)$$

in the same space as ρ . In order to distinguish it from the real system trajectory, let us denote $\hat{\mathbf{s}}_{0,1,\dots}^\pi$ the (possibly stochastic) trajectories of the state transition model (7) under policy π , i.e., $\hat{\mathbf{s}}_{k+1}^\pi \sim \hat{\rho}[\cdot | \hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)]$, starting from $\hat{\mathbf{s}}_0^\pi = \mathbf{s}, \forall \pi$. We further denote the measure associated with such trajectories as $\hat{\tau}^\pi$. In general, $\hat{\cdot}$ refers to the notations related to the imperfect model of the system in this paper. It has been shown in [17] that proving closed-loop stability of the Markov Chains with the optimal policy resulting from an undiscounted OCP is more straightforward than a discounted setting [15]. This observation is well-known in MPC of deterministic systems [18]. Therefore, in this paper, we are interested in using an undiscounted OCP for the model (7) in order to extract the optimal policy and optimal value functions of the real system (1), as this allows us to enforce stability guarantees.

A. Finite-horizon OCP

While MPC allows one to introduce stability and safety guarantees, it also requires a model of the real system which is bound to be imperfect, and it optimizes the cost over a finite horizon with unitary discount factor. In other words, MPC is an MDP based on the imperfect system model (7) which we will formulate in (8). In this section we will prove that these differences between the MPC formulation and the original MDP formulation do not hinder the ability to obtain the optimal policy and the optimal value functions of the real system through MPC. Consider the following undiscounted finite-horizon OCP associated to model (7):

$$\hat{V}_N^*(\mathbf{s}) = \min_{\pi} \hat{V}_N^\pi(\mathbf{s}) := \mathbb{E}_{\hat{\tau}^\pi} \left[\hat{T}(\hat{\mathbf{s}}_N^\pi) + \sum_{k=0}^{N-1} \hat{L}(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) \right], \quad (8)$$

with initial state $\hat{\mathbf{s}}_0^\pi = \mathbf{s}$, where $N \in \mathbb{N}$ is the horizon length, \hat{T} , \hat{L} , \hat{V}_N^* and \hat{V}_N^π are the terminal cost, the stage cost, the optimal

value function and the value function of the policy π associated to model (7), respectively, and where \mathbb{N} is the set of natural numbers. The expectation $\mathbb{E}_{\hat{\tau}\pi}$ in (8) is taken over undiscounted closed-loop Markov Chain (7) with policy π . We denote $\hat{\pi}_N^*$ the optimal policy resulting from (8). Moreover, the action-value function \hat{Q}_N^* associated to (8) is defined as follows:

$$\hat{Q}_N^*(\mathbf{s}, \mathbf{a}) := \hat{L}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\hat{\rho}} \left[\hat{V}_{N-1}^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a} \right], \quad (9a)$$

$$\hat{V}_0^*(\mathbf{s}) := \hat{T}(\mathbf{s}) \quad (9b)$$

The next assumption expresses a requirement on the boundedness of V^* under model trajectories $\hat{\mathbf{s}}_{0,1,\dots}^\pi$ with the optimal policy π^* which allows us to develop the theoretical results of this paper.

Assumption 1. *The following set is non-empty for a given $\bar{N} \in \mathbb{N}$.*

$$\mathcal{S} := \left\{ \mathbf{s} \in \mathcal{X} \mid \left| \mathbb{E}_{\hat{\tau}\pi^*} \left[V^*(\hat{\mathbf{s}}_k^{\pi^*}) \right] \right| < \infty, \forall k \leq \bar{N} \right\} \quad (10)$$

Assumption 1 requires that there exists a non-empty set \mathcal{S} such that for all trajectories starting in it, the expected value of V^* is bounded at all future times under the state distribution given by the model in finite time under the optimal policy. This assumption plays a vital role in the derivation of our main result. We will further detail this assumption in Remark 1.

The next theorem provides theoretical support to the idea that one can recover the optimal policy and value functions by means of an MPC scheme which is based on an imperfect model and has an undiscounted formulation over a finite prediction horizon.

Theorem 1. *Suppose that Assumption 1 holds for $\bar{N} \geq N$. Then, there exist a terminal cost \hat{T} and a stage cost \hat{L} such that the following identities hold, $\forall \gamma, N \in \mathbb{N}$ and $\mathbf{s} \in \mathcal{S}$:*

- (i) $\hat{\pi}_N^*(\mathbf{s}) = \pi^*(\mathbf{s})$,
- (ii) $\hat{V}_N^*(\mathbf{s}) = V^*(\mathbf{s})$,
- (iii) $\hat{Q}_N^*(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a})$, for the inputs $\mathbf{a} \in \mathcal{U}$ such that $|\mathbb{E}_{\hat{\rho}} [V^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}]| < \infty$

Proof. We select the terminal cost \hat{T} and the stage cost \hat{L} as follows:

$$\hat{T}(\mathbf{s}) = V^*(\mathbf{s}) \quad (11a)$$

$$\hat{L}(\mathbf{s}, \mathbf{a}) = \begin{cases} Q^*(\mathbf{s}, \mathbf{a}) - \mathbb{E}_{\hat{\rho}} [V^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}] & \text{If } |\mathbb{E}_{\hat{\rho}} [V^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}]| < \infty \\ \infty & \text{otherwise} \end{cases} \quad (11b)$$

Under Assumption 1, the terminal and stage costs in (8) have a finite expected value for all $\hat{\mathbf{s}}_0^{\pi^*} \in \mathcal{S}$. By substitution of (11) in (8) and using telescopic sum, we have:

$$\begin{aligned} \hat{V}_N^*(\mathbf{s}) &= \mathbb{E}_{\hat{\tau}\pi} \left[\hat{T}(\hat{\mathbf{s}}_N^\pi) + \sum_{k=0}^{N-1} \hat{L}(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) \right] \\ &\stackrel{(11)}{=} \mathbb{E}_{\hat{\tau}\pi} \left[V^*(\hat{\mathbf{s}}_N^\pi) + \sum_{k=0}^{N-1} \left(Q^*(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) - V^*(\hat{\mathbf{s}}_{k+1}^\pi) \right) \right] \\ &= Q^*(\mathbf{s}, \pi(\mathbf{s})) + \mathbb{E}_{\hat{\tau}\pi} \left[\sum_{k=1}^{N-1} \left(Q^*(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) - V^*(\hat{\mathbf{s}}_k^\pi) \right) \right] \\ &= Q^*(\mathbf{s}, \pi(\mathbf{s})) + \mathbb{E}_{\hat{\tau}\pi} \left[\sum_{k=1}^{N-1} A^*(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) \right], \quad (12) \end{aligned}$$

where $\hat{\mathbf{s}}_0 = \mathbf{s}$. From (4a) and (4b), we know that:

$$\pi^*(\cdot) = \arg \min_{\pi} A^*(\cdot, \pi(\cdot)) = \arg \min_{\pi} Q^*(\cdot, \pi(\cdot)) \quad (13)$$

then from (12):

$$\begin{aligned} \pi^*(\mathbf{s}) &= \arg \min_{\pi} \hat{V}_N^*(\mathbf{s}) \\ &= \arg \min_{\pi} Q^*(\mathbf{s}, \pi(\mathbf{s})) + \mathbb{E}_{\hat{\tau}\pi} \left[\sum_{k=1}^{N-1} A^*(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) \right] \end{aligned} \quad (14)$$

Note that π^* minimizes all terms in the cost above, i.e., A^* and Q^* , such that it must also minimize \hat{V}_N^* . This proves (i), i.e.,

$$\pi^*(\mathbf{s}) = \hat{\pi}_N^*(\mathbf{s}).$$

In turn, this proves (ii), since

$$\begin{aligned} \hat{V}_N^*(\mathbf{s}) &= \hat{V}_N^{\pi^*}(\mathbf{s}) = Q^*(\mathbf{s}, \pi^*(\mathbf{s})) + \\ &+ \mathbb{E}_{\hat{\tau}\pi^*} \left[\underbrace{\sum_{k=1}^N A^*(\hat{\mathbf{s}}_k^{\pi^*}, \pi^*(\hat{\mathbf{s}}_k^{\pi^*}))}_{\stackrel{(4b)}{=} 0} \mid \hat{\mathbf{s}}_0 = \mathbf{s} \right] \\ &= Q^*(\mathbf{s}, \pi^*(\mathbf{s})) \stackrel{(4a)}{=} V^*(\mathbf{s}). \quad (15) \end{aligned}$$

Moreover, from (9a) and (11b), for any inputs $\mathbf{a} \in \mathcal{U}$ such that $|\mathbb{E}_{\hat{\rho}} [V^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}]| < \infty$, we have:

$$\begin{aligned} \hat{Q}_N^*(\mathbf{s}, \mathbf{a}) &= \hat{L}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\hat{\rho}} \left[\hat{V}_{N-1}^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a} \right] \\ &\stackrel{(11b)}{=} Q^*(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\hat{\rho}} \left[\hat{V}_{N-1}^*(\mathbf{s}^+) - V^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a} \right] = Q^*(\mathbf{s}, \mathbf{a}), \end{aligned} \quad (16)$$

where the last inequality is obtained by noting that (ii) for $N > 1$ and $\hat{V}_0^*(\mathbf{s}) = \hat{T}(\mathbf{s}) = V^*(\mathbf{s})$ for $N = 1$. This directly yields (iii). ■

Theorem 1 states that, independent of the discount factor γ , it is possible to find a finite-horizon OCP cost function that provides the optimal policy and optimal value functions of a discounted MDP if an inexact model is used in the finite-horizon OCP. We observe that the setup of this paper has been analyzed in [15], under the assumption of a perfect model, i.e., $\hat{\rho}[\cdot | \mathbf{s}, \mathbf{a}] = \rho[\cdot | \mathbf{s}, \mathbf{a}]$. In that case (11b) reads:

$$\hat{L}(\mathbf{s}, \mathbf{a}) = \ell(\mathbf{s}, \mathbf{a}) + (\gamma - 1) \mathbb{E}_{\rho} [V^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}], \quad \forall \mathbf{s} \in \mathcal{S}, \quad (17)$$

which corresponds to the cost modification discussed in [15].

B. Infinite-horizon OCP

In this section, we investigate the case $N \rightarrow \infty$ for which, under some conditions, the terminal cost can be dismissed. In this case, we first make the next additional assumption.

Assumption 2. *We assume that the optimal value function converges to a constant and finite value with model (7) under the optimal policy π^* . I.e.:*

$$-\infty < \lim_{N \rightarrow \infty} \mathbb{E}_{\hat{\tau}\pi^*} \left[V^*(\hat{\mathbf{s}}_N^{\pi^*}) \right] = \hat{v}_\infty < \infty \quad (18)$$

Assumption 2 can be interpreted as some forms of the stability condition on the model dynamics under the optimal policy π^* . We will explain this assumption in Remark 1. In this section, we consider the following undiscounted value function without terminal cost:

$$\hat{V}_\infty^*(\mathbf{s}) := \min_{\pi} \hat{V}_\infty^\pi(\mathbf{s}) := \lim_{N \rightarrow \infty} \mathbb{E}_{\hat{\tau}\pi} \left[\sum_{k=0}^{N-1} \hat{L}(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) \right] \quad (19)$$

with initial state $\hat{\mathbf{s}}_0^\pi = \mathbf{s}$. We denote the optimal policy solution of (19) as $\hat{\pi}_\infty^*(\mathbf{s})$. We then define the optimal action-value function \hat{Q}_∞^* associated to (19) as follows:

$$\hat{Q}_\infty^*(\mathbf{s}, \mathbf{a}) = \hat{L}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\hat{\rho}} \left[\hat{V}_\infty^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a} \right], \quad (20)$$

We are now ready to state the equivalent of Theorem 1 in case of an infinite horizon without a terminal cost.

Theorem 2. *Suppose that Assumptions 1 and 2 hold, then the following hold $\forall \mathbf{s} \in \mathcal{S}, \forall \gamma$:*

- (i) $\hat{\pi}_\infty^*(\mathbf{s}) = \pi^*(\mathbf{s})$
- (ii) $\hat{V}_\infty^*(\mathbf{s}) = V^*(\mathbf{s}) - \hat{v}_\infty$
- (iii) $\hat{Q}_\infty^*(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a}) - \hat{v}_\infty$, for the inputs $\mathbf{a} \in \mathcal{U}$ such that $|\mathbb{E}_{\hat{\rho}}[V^*(\mathbf{s}^+)|\mathbf{s}, \mathbf{a}]| < \infty$

if the stage cost \hat{L} is selected according Equation (11b).

Proof. Using stage cost \hat{L} in (11b), we have:

$$\begin{aligned} \hat{V}_\infty^\pi(\mathbf{s}) &= \lim_{N \rightarrow \infty} \mathbb{E}_{\hat{\tau}\pi} \left[\sum_{k=0}^{N-1} Q^*(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) - \right. \\ &\quad \left. \mathbb{E}_{\hat{\rho}} [V^*(\hat{\mathbf{s}}_{k+1}^\pi) | \hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)] \right] \\ &= \lim_{N \rightarrow \infty} \mathbb{E}_{\hat{\tau}\pi} \left[\sum_{k=0}^{N-1} Q^*(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) - V^*(\hat{\mathbf{s}}_{k+1}^\pi) \right] \\ &= Q^*(\mathbf{s}, \pi(\mathbf{s})) + \lim_{N \rightarrow \infty} \mathbb{E}_{\hat{\tau}\pi} \left[-V^*(\hat{\mathbf{s}}_N^\pi) + \right. \\ &\quad \left. \sum_{k=1}^{N-1} Q^*(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) - V^*(\hat{\mathbf{s}}_k^\pi) \right] \\ &= Q^*(\mathbf{s}, \pi(\mathbf{s})) + \lim_{N \rightarrow \infty} \mathbb{E}_{\hat{\tau}\pi} \left[-V^*(\hat{\mathbf{s}}_N^\pi) + \sum_{k=1}^{N-1} A^*(\hat{\mathbf{s}}_k^\pi, \pi(\hat{\mathbf{s}}_k^\pi)) \right] \end{aligned} \quad (21)$$

where $\hat{\mathbf{s}}_0^\pi = \mathbf{s}$. By (4a) and (4b) we know that the policy $\pi(\mathbf{s}) = \pi^*(\mathbf{s})$ minimizes all terms $A^*(\cdot, \pi(\cdot))$ and $Q^*(\cdot, \pi(\cdot))$, such that it also minimizes $\hat{V}_\infty^\pi(\mathbf{s})$, i.e.:

$$\hat{\pi}_\infty^*(\mathbf{s}) = \arg \min_{\pi} \hat{V}_\infty^\pi(\mathbf{s}) = \pi^*(\mathbf{s}), \quad (22)$$

which proves (i). Moreover:

$$\hat{V}_\infty^{\pi^*}(\mathbf{s}) = V^*(\mathbf{s}) - \lim_{N \rightarrow \infty} \mathbb{E} [V^*(\hat{\mathbf{s}}_N^{\pi^*})]. \quad (23)$$

Using (18) we have:

$$\hat{V}_\infty^*(\mathbf{s}) = \hat{V}_\infty^{\pi^*}(\mathbf{s}) = V^*(\mathbf{s}) - \hat{v}_\infty. \quad (24)$$

For the inputs $\mathbf{a} \in \mathcal{U}$ such that $|\mathbb{E}_{\hat{\rho}}[V^*(\mathbf{s}^+)|\mathbf{s}, \mathbf{a}]| < \infty$:

$$\begin{aligned} \hat{Q}_\infty^*(\mathbf{s}, \mathbf{a}) &= \hat{L}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\hat{\rho}} [\hat{V}_\infty^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}] \\ &= Q^*(\mathbf{s}, \mathbf{a}) - \mathbb{E}_{\hat{\rho}} [V^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}] + \mathbb{E}_{\hat{\rho}} [\hat{V}_\infty^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}] \\ &= Q^*(\mathbf{s}, \mathbf{a}) - \mathbb{E}_{\hat{\rho}} [V_\infty^*(\mathbf{s}^+) - \hat{V}_\infty^*(\mathbf{s}^+) | \mathbf{s}, \mathbf{a}] = Q^*(\mathbf{s}, \mathbf{a}) - \hat{v}_\infty, \end{aligned} \quad (25)$$

which completes the proof. \blacksquare

Theorem 2 extends Theorem 1 to the case of an infinite horizon with zero terminal cost. Assumption 2 is necessary in order to be able to remove the terminal cost. In the next section we will detail the use of the theorems in practice and reformulate OCP (8) as a Model Predictive Control (MPC) scheme.

IV. MPC AS A FUNCTION APPROXIMATOR FOR RL

As it was shown in the previous section, the optimal policy and value functions of any MDP with either discounted or undiscounted criteria can be captured using a finite-horizon undiscounted OCP (8) even if the model is not accurate. Since the equivalence only holds at the initial state, if one is interested in recovering the optimal MDP policy, the finite-horizon OCP needs to be solved from scratch for each

initial state. In practice, this amounts to deploying the finite-horizon OCP in an MPC framework, i.e., in a closed-loop.

As discussed above, the equivalence is only obtained if a properly modified stage and terminal costs are introduced for the finite-horizon undiscounted MPC scheme. However, finding such costs requires knowledge about the optimal value functions of the real MDP. In this section, we detail how the theorems we provided in the previous sections can be used in practice to exploit MPC as a structured function approximator of the optimal policy and value functions of the real MDP. One of the main advantages of MPC is that it allows us to straightforwardly introduce state and input constraints in the policy. We parameterize the MPC scheme with parameter vector θ such that RL methods can be deployed to tune θ in order to achieve the equivalence yielding the optimal policy and value functions of the real system and, consequently, the best possible closed-loop performance.

As the MPC model is not required to capture the real system dynamics exactly, for the sake of reducing the computational burden, and due to the (relative) simplicity of the resulting MPC scheme, a popular choice of model $\hat{\rho}[\mathbf{s}^+|\mathbf{s}, \mathbf{a}]$ is a deterministic model, i.e.:

$$\hat{\rho}[\mathbf{s}^+|\mathbf{s}, \mathbf{a}] = \delta(\mathbf{s}^+ - \mathbf{f}_\theta(\mathbf{s}, \mathbf{a})) \quad (26)$$

where $\delta(\cdot)$ is the Dirac measure and $\mathbf{f}_\theta(\mathbf{s}, \mathbf{a})$ is a parameterized deterministic (possibly nonlinear) model. We approximate the modified costs \hat{L} and \hat{T} by parametric functions L_θ and T_θ , respectively. Due to the mismatch between the model and the real system, hard constraints in the MPC scheme could become infeasible. This is a well-known issue in the MPC community and one simple solution consists in formulating the state constraints as soft constraints [19]. We therefore formulate the MPC finite-horizon OCP as:

$$\begin{aligned} \hat{V}_N^\theta(\mathbf{s}) &= \min_{\hat{\mathbf{a}}, \hat{\mathbf{s}}, \boldsymbol{\sigma}} -\lambda_\theta(\hat{\mathbf{s}}_0) + T_\theta(\hat{\mathbf{s}}_N) + \boldsymbol{\mu}_f^\top \boldsymbol{\sigma}_N \\ &\quad + \sum_{k=0}^{N-1} L_\theta(\hat{\mathbf{s}}_k, \hat{\mathbf{a}}_k) + \boldsymbol{\mu}^\top \boldsymbol{\sigma}_k \end{aligned} \quad (27a)$$

$$\text{s.t. } \hat{\mathbf{s}}_{k+1} = \mathbf{f}_\theta(\hat{\mathbf{s}}_k, \hat{\mathbf{a}}_k), \quad \hat{\mathbf{s}}_0 = \mathbf{s}, \quad (27b)$$

$$\hat{\mathbf{a}}_k \in \mathcal{U}, \quad 0 \leq \boldsymbol{\sigma}_k, \quad 0 \leq \boldsymbol{\sigma}_N, \quad (27c)$$

$$\mathbf{h}_\theta(\hat{\mathbf{s}}_k, \hat{\mathbf{a}}_k) \leq \boldsymbol{\sigma}_k^*, \quad \mathbf{h}_\theta^f(\hat{\mathbf{s}}_N) \leq \boldsymbol{\sigma}_N^*, \quad (27d)$$

where \hat{V}_N^θ is the MPC-based parameterized value function, $\mathbf{h}_\theta(\mathbf{s}, \mathbf{a})$ is a mixed input-state constraint, $\mathbf{h}_\theta^f(\mathbf{s})$ is the terminal constraint, $\boldsymbol{\sigma}_k$ and $\boldsymbol{\sigma}_N$ are slack variables guaranteeing the feasibility of the MPC scheme and $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_f$ are constant vectors that ought to be selected sufficiently large [19]. Note that these constants allow the MPC scheme to find a feasible solution, but penalize constraint violations enough to guarantee that a feasible solution is found whenever possible. While alternative feasibility-enforcing strategies, e.g., robust MPC, do exist, an exhaustive discussion on the topic is beyond the scope of this paper. Function λ_θ parameterizes the so-called storage function, which has been added to the cost in order to enable the MPC scheme to tackle the case of so-called economic problems. Such situations arise when the MDP stage cost is not positive definite, while the MPC stage cost is forced to be positive definite in order to obtain a stabilizing feedback policy. Note that since the term $-\lambda_\theta(\hat{\mathbf{s}}_0)$ only depends on the current state, it does not modify the optimal policy. For more details, we refer the interested readers to [9], [20].

While Theorem 1 states that one can find suitable stage and terminal costs for any given model, adjusting the model parameters is not essential from the theoretical perspective. However, in practice, the stage and the terminal cost parameterization may not capture \hat{L} and \hat{T} exactly. Since \hat{L} and \hat{T} are (implicitly) functions of the model, using a parameterized model \mathbf{f}_θ introduces extra degrees of freedom to bring \hat{L} and \hat{T} closer to the functions that can be represented by L_θ

and T_θ . In turn, this can yield a better approximation of the optimal policy and value function. The MPC parameterized policy can be obtained from (27) as follows:

$$\hat{\pi}_N^\theta(\mathbf{s}) = \hat{\mathbf{a}}_0^*(\theta, \mathbf{s}), \quad (28)$$

where $\hat{\mathbf{a}}_0^*$ is the solution of (27), corresponding to the first input $\hat{\mathbf{a}}_0$. Moreover, the parameterized action-value function based on MPC scheme (27) can be formulated as follows:

$$\hat{Q}_N^\theta(\mathbf{s}, \mathbf{a}) := \min_{\hat{\mathbf{a}}, \hat{\mathbf{s}}, \hat{\sigma}} (27a), \quad \text{s.t. (27b) - (27d)}, \quad \hat{\mathbf{a}}_0 = \mathbf{a}. \quad (29)$$

Then one obtains the following identities:

$$\hat{V}_N^\theta(\mathbf{s}) = \min_{\hat{\mathbf{a}}} \hat{Q}_N^\theta(\mathbf{s}, \hat{\mathbf{a}}), \quad \hat{\pi}_N^\theta(\mathbf{s}) \in \arg \min_{\hat{\mathbf{a}}} \hat{Q}_N^\theta(\mathbf{s}, \hat{\mathbf{a}}). \quad (30)$$

We can use RL techniques, such as Q-learning and policy gradient method to tune the parameters θ of parameterized MPC scheme (27) and approach the *optimal* parameter θ^* . For instance, at each learning step, Q-learning based on Temporal difference (TD) method uses the following update rule for θ :

$$\delta_k := \ell(\mathbf{s}_k, \mathbf{a}_k) + \gamma \hat{V}_N^\theta(\mathbf{s}_{k+1}) - \hat{Q}_N^\theta(\mathbf{s}_k, \mathbf{a}_k) \quad (31a)$$

$$\theta \leftarrow \theta + \zeta \delta_k \nabla_{\theta} \hat{Q}_N^\theta(\mathbf{s}_k, \mathbf{a}_k) \quad (31b)$$

in order to capture the optimal value function $\hat{Q}_N^{\theta^*} \approx Q^*$ for the optimal parameters θ^* , where the scalar $\zeta > 0$ is the learning step-size and δ_k is the TD error. The use of RL for the tuning the MPC scheme can be found e.g., in [9], [21].

V. ANALYTICAL CASE STUDY

We consider a Linear Quadratic Regulator (LQR) example in order to obtain the corresponding optimal value functions analytically and verify Theorem 2. The real system and stage cost are given as follows:

$$\mathbf{s}^+ = \mathbf{A}\mathbf{s} + \mathbf{B}\mathbf{a} + \mathbf{e}, \quad \ell(\mathbf{s}, \mathbf{a}) = \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix}^\top \begin{bmatrix} T & N \\ N^\top & R \end{bmatrix} \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix}, \quad (32)$$

where $\mathbf{e} \sim \mathcal{N}(0, \Sigma)$ with the discount factor γ . One can verify the following optimal value functions:

$$V^*(\mathbf{s}) = \mathbf{s}^\top S \mathbf{s} + \hat{v}_\infty, \quad (33)$$

$$Q^*(\mathbf{s}, \mathbf{a}) = \hat{v}_\infty + \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix}^\top \begin{bmatrix} T + \gamma \mathbf{A}^\top S \mathbf{A} & N + \gamma \mathbf{A}^\top S \mathbf{B} \\ N^\top + \gamma \mathbf{B}^\top S \mathbf{A} & R + \gamma \mathbf{B}^\top S \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix},$$

where $\hat{v}_\infty = \frac{\gamma}{1-\gamma} \text{Tr}(S\Sigma)$ and S is obtained from the following Riccati equations:

$$T + \gamma \mathbf{A}^\top S \mathbf{A} = S + (N + \gamma \mathbf{A}^\top S \mathbf{B}) (K_\gamma^*)^\top, \quad (34a)$$

$$(R + \gamma \mathbf{B}^\top S \mathbf{B}) K_\gamma^* = N^\top + \gamma \mathbf{B}^\top S \mathbf{A}. \quad (34b)$$

Then $\pi^*(\mathbf{s}) = -K_\gamma^* \mathbf{s}$ and $\bar{\pi}^*(\mathbf{s}) = \tilde{\pi}^*(\mathbf{s}) = -K_1^* \mathbf{s}$, where $K_1^* = \lim_{\gamma \rightarrow 1} K_\gamma^*$. We then consider a linear deterministic model:

$$\mathbf{s}^+ = \hat{\mathbf{A}}\mathbf{s} + \hat{\mathbf{B}}\mathbf{a}, \quad (35)$$

and an undiscounted OCP with the following stage cost, defined according to Equation (11b) as:

$$\hat{\ell}(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a}) - V^*(\hat{\mathbf{s}}^+) \quad (36)$$

$$\stackrel{(33)}{=} \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix}^\top \begin{bmatrix} T + \gamma \mathbf{A}^\top S \mathbf{A} & N + \gamma \mathbf{A}^\top S \mathbf{B} \\ N^\top + \gamma \mathbf{B}^\top S \mathbf{A} & R + \gamma \mathbf{B}^\top S \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix} \\ - (\hat{\mathbf{A}}\mathbf{s} + \hat{\mathbf{B}}\mathbf{a})^\top S (\hat{\mathbf{A}}\mathbf{s} + \hat{\mathbf{B}}\mathbf{a}) := \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix}^\top \begin{bmatrix} \hat{T} & \hat{N} \\ \hat{N}^\top & \hat{R} \end{bmatrix} \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix}.$$

The Riccati equations for the undiscounted problem with the model (35) read as:

$$\hat{T} + \hat{\mathbf{A}}^\top \hat{S} \hat{\mathbf{A}} = \hat{S} + (\hat{N} + \hat{\mathbf{A}}^\top \hat{S} \hat{\mathbf{B}}) (\hat{K}^*)^\top, \quad (37a)$$

$$(\hat{R} + \hat{\mathbf{B}}^\top \hat{S} \hat{\mathbf{B}}) \hat{K}^* = \hat{N}^\top + \hat{\mathbf{B}}^\top \hat{S} \hat{\mathbf{A}}. \quad (37b)$$

with the optimal policy $\hat{\pi}^*(\mathbf{s}) = -\hat{K}^* \mathbf{s}$ and the optimal value function $\hat{V}_\infty^*(\mathbf{s}) = \mathbf{s}^\top \hat{S} \mathbf{s}$. From (36), we have:

$$T + \gamma \mathbf{A}^\top S \mathbf{A} - \hat{\mathbf{A}}^\top S \hat{\mathbf{A}} = \hat{T}, \quad (38a)$$

$$N + \gamma \mathbf{A}^\top S \mathbf{B} - \hat{\mathbf{A}}^\top S \hat{\mathbf{B}} = \hat{N}, \quad (38b)$$

$$R + \gamma \mathbf{B}^\top S \mathbf{B} - \hat{\mathbf{B}}^\top S \hat{\mathbf{B}} = \hat{R}. \quad (38c)$$

Equivalently, this entails that \hat{T} , \hat{N} and \hat{R} must satisfy

$$\hat{T} + \hat{\mathbf{A}}^\top S \hat{\mathbf{A}} = T + \gamma \mathbf{A}^\top S \mathbf{A}, \quad (39a)$$

$$\hat{N} + \hat{\mathbf{A}}^\top S \hat{\mathbf{B}} = N + \gamma \mathbf{A}^\top S \mathbf{B}, \quad (39b)$$

$$\hat{R} + \hat{\mathbf{B}}^\top S \hat{\mathbf{B}} = R + \gamma \mathbf{B}^\top S \mathbf{B}. \quad (39c)$$

Then:

$$\hat{T} + \hat{\mathbf{A}}^\top S \hat{\mathbf{A}} \stackrel{(39a)}{=} T + \gamma \mathbf{A}^\top S \mathbf{A} \stackrel{(34a)}{=} S + \\ S(N + \gamma \mathbf{A}^\top S \mathbf{B}) (K_\gamma^*)^\top \stackrel{(39b)}{=} S + (\hat{N} + \hat{\mathbf{A}}^\top S \hat{\mathbf{B}}) (K_\gamma^*)^\top, \quad (40)$$

and

$$(\hat{R} + \hat{\mathbf{B}}^\top S \hat{\mathbf{B}}) K_\gamma^* \stackrel{(39c)}{=} (R + \gamma \mathbf{B}^\top S \mathbf{B}) K_\gamma^* \\ \stackrel{(34b)}{=} N^\top + \gamma \mathbf{B}^\top S \mathbf{A} \stackrel{(39b)}{=} \hat{N}^\top + \hat{\mathbf{B}}^\top S \hat{\mathbf{B}}. \quad (41)$$

Equations (40) and (41) show that $\hat{S} = S$ and $\hat{K}^* = K_\gamma^*$ satisfy the undiscounted Riccati equations (37). Then it reads that $\pi^*(\mathbf{s}) = \hat{\pi}^*(\mathbf{s})$ and $V^*(\mathbf{s}) = \hat{V}_\infty^*(\mathbf{s}) + \hat{v}_\infty$.

Remark 1. Regarding Assumption 1, the value function will remain bounded in the finite horizon prediction for every bounded initial condition \mathbf{s}_0 and every linear model in form (35) for a given control policy $\pi^*(\mathbf{s}) = -K_\gamma^* \mathbf{s}$ or $\bar{\pi}^*(\mathbf{s}) = \tilde{\pi}^*(\mathbf{s}) = -K_1^* \mathbf{s}$. For Assumption 2, the linear model matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ must be chosen such that $\rho(\hat{\mathbf{A}} - \hat{\mathbf{B}} K_\gamma^*) \leq 1$ in order to guarantee boundedness of the optimal value function (33). For instance, for a scalar dynamics, the locus of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ is shown in Figure 1. Inspired by this example, we ought to point out here that for linear systems Assumption 1 is automatically obtained if the model is stabilized by the optimal policy, though the converse might not be true (e.g., if the cost is 0). Note that, the systems without constraint satisfying Assumption 1 is fairly straightforward while in the presence of the system constraints, the model also must not violate those constraints. To satisfy Assumption 2, a model must be adopted whose trajectory does not diverge under the optimal policy of the real system and satisfy the system constraint. It is clear that the closer the model is to the real system the more likely it is to satisfy this assumption. This model can be obtained based on offline system identification. In [22], the authors proposed to use robust MPC in order to ensure constraint satisfaction. A deeper discussion of these assumptions can be found in [9] and [15].

VI. NUMERICAL EXAMPLES

In this section, we will investigate some numerical example to illustrate the theoretical development.

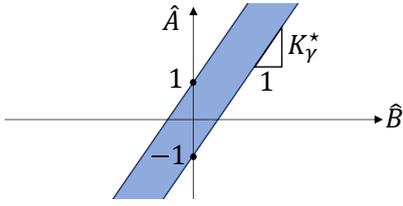


Fig. 1. The blue area shows all \hat{A} and \hat{B} in the linear model such that the resulting trajectory and optimal value function remain bounded for the given optimal policy $\pi^*(s) = -K_Y^*s$.

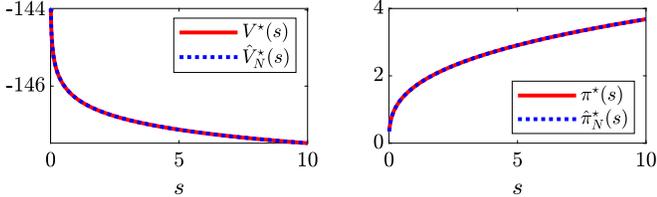


Fig. 2. (Left:) Optimal value functions (Right:) and optimal policy resulting from the discounted real system and undiscounted MPC scheme with the wrong model.

A. Non-quadratic stage cost

In this example, we provide a benchmark optimal investment problem with a non-quadratic stage cost. Consider the following dynamics and stage cost [23]:

$$s_{k+1} = a_k, \quad \ell(s, a) = -\ln(As^\alpha - a), \quad (42)$$

where A and $0 < \alpha < 1$ are given constants. It is known that for the discount factor γ , the optimal value and policy functions are $V^*(s) = B + C \ln(s)$ and $\pi^*(s) = \gamma\alpha As^\alpha$, where [24]:

$$B = \frac{\ln((1-\alpha\gamma)A) + \frac{\gamma\alpha}{1-\gamma\alpha} \ln(\alpha\gamma A)}{\gamma-1}, \quad C = \frac{\alpha}{\alpha\gamma-1}. \quad (43)$$

We then consider a model of the dynamics with $\hat{s}_{k+1} = \mu\hat{a}_k$ and, based on this model, we construct a finite-horizon undiscounted MPC with the costs according Equation (11) in Theorem 1 and $N = 10$. In this example we have considered $A = 5$, $\alpha = 0.34$, $\mu = 0.8$ and $\gamma = 0.9$. Figure 2 compares the optimal value and policy functions from the discounted real system (42) and from the MPC scheme with a wrong model. As predicted by Theorem 1, one can see that they match perfectly. Note that the results are valid for every discount factor $0 < \gamma < 1$, every horizon length and for other values of the constants A , α , and μ .

B. Inverted pendulum with process noise

We consider the following discrete-time stochastic dynamics, representing an inverted pendulum with a random support excitation:

$$s_{k+1} = s_k + \begin{bmatrix} s_k(2) \\ (\frac{g}{l} + \xi) \sin(s_k(1)) \end{bmatrix} \delta t + \begin{bmatrix} 0 \\ \frac{\delta t}{ml^2} \end{bmatrix} a_k \quad (44)$$

where $g = 9.81$, $l = 0.3$, $m = 0.5$ and $\delta t = 0.1$ are constants representing the gravity, mass, length and the sampling time of the discrete dynamics. Disturbance $\xi \sim \mathcal{U}[-0.5, 0.5]$ has a uniform distribution and $s_k := [s_k(1), s_k(2)]^\top$ is the system state and a_k is the system input. We consider $\ell(s, a) = s^\top s + a^2$ as a stage cost with the discount factor $\gamma = 0.95$. We first aim to find an approximate solution for the optimal policy and the optimal value functions using Dynamic Programming (DP). We consider the state constraints $-1 \leq s_k(1) \leq 1$, $-1 \leq s_k(2) \leq 1$ and the input constraint $-0.8 \leq a_k \leq 0.8$. Figure 3 shows the optimal value

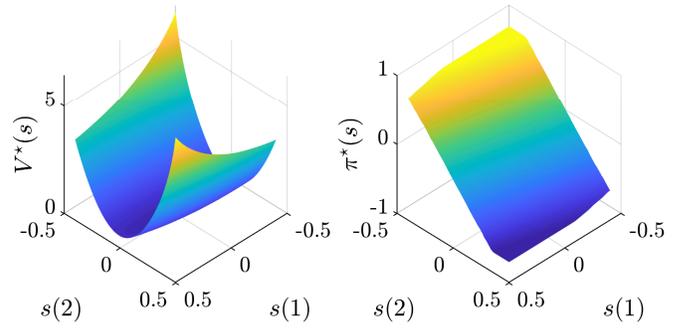


Fig. 3. Optimal Value (left) and policy (right) functions resulting from ADP.

function and the optimal policy function resulting from DP for the discounted infinite-horizon MDP.

We build an undiscounted finite-horizon OCP with a wrong model in order to capture the optimal value and the optimal policy functions of the discounted infinite horizon MDP. To do this, we consider an MPC scheme with a deterministic linearized form of the dynamics as a model of the real system as follows:

$$\hat{s}_{k+1} = f_\theta(\hat{s}_k, \hat{a}_k) = \hat{s}_k + \begin{bmatrix} \hat{s}_k(2) \\ \frac{g}{\theta_l} \hat{s}_k(1) \end{bmatrix} \delta t + \begin{bmatrix} 0 \\ \frac{\delta t}{m\theta_l^2} \end{bmatrix} \hat{a}_k \quad (45)$$

where $\hat{s}_k := [\hat{s}_k(1), \hat{s}_k(2)]^\top$ and \hat{a}_k are the model state and input. Moreover, we consider an uncertain l with a adjustable parameter θ_l , with an initial value 0.25. We consider the parameterized MPC scheme with the horizon length $N = 10$ and the following parameterized quadratic stage and terminal cost:

$$T_\theta(s) = s^\top G s, \quad L_\theta(s, a) = \begin{bmatrix} s \\ a \end{bmatrix}^\top H \begin{bmatrix} s \\ a \end{bmatrix} \quad (46)$$

where G and H are parametric positive definite matrices. Then the parameters vector θ gathers all the adjustable parameters as $\theta = \{\theta_l, G, H\}$. We use the Q-learning method in order to update the parameters θ to achieve the optimal solutions of the real system and improve the closed-loop performance. Figure 4 shows the difference between the MPC value \hat{V}_N^θ and policy $\hat{\pi}_N^\theta$ functions with their optimal solutions computed by DP. The blue and red surfaces represent this difference at the beginning of the learning and after 500 learning steps, respectively. As it can be seen, the results are getting closer to zero as the learning proceeds. Note that the stage and terminal costs yielding a perfect match of V^* and π^* , as per Theorem 1, do not have a quadratic form, hence the selected MPC formulation cannot capture them exactly. The green surfaces in Figure 4 have been obtained by computing these stage and terminal costs numerically and shows the corresponding $\hat{V}_N^* - V^*$ and $\hat{\pi}_N^* - \pi^*$. As expected the difference is zero, modulo tiny numerical inaccuracies.

Finally, Figure 5 illustrates the closed-loop performance of the system under the MPC policy $\hat{\pi}_N^\theta$. As the closed loop cost decreases, this demonstrates that RL can be effective in tuning the MPC parameters so as to achieve the best closed-loop performance.

C. Learning based MPC: Tracking stage cost

In this section, we consider the cart-pendulum balancing problem shown in Figure 6 in order to illustrate the proposed method in a constrained tracking problem. The dynamics are given by:

$$(M + m)\ddot{x} + \frac{1}{2}ml\ddot{\phi} \cos \phi = \frac{1}{2}ml\dot{\phi}^2 \sin \phi + u, \quad (47a)$$

$$\frac{1}{3}ml^2\ddot{\phi} + \frac{1}{2}ml\ddot{x} \cos \phi = -\frac{1}{2}mgl \sin \phi, \quad (47b)$$

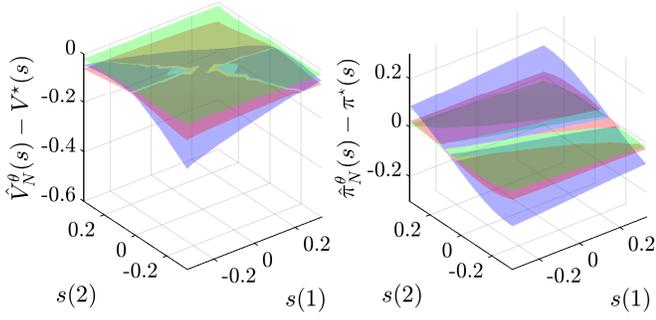


Fig. 4. The difference between the MPC based parameterized value (left) policy (right) and their optimal solutions for the beginning of the learning (blue) and after 500 learning steps (red) and the exact cost modification from theorem 1 (green).

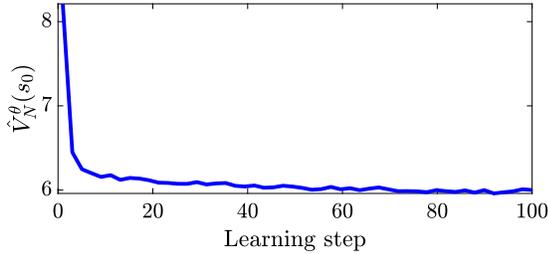


Fig. 5. The MPC-based value function $\hat{V}_N^\theta(s_0)$ during the learning.

where M and m are the cart mass and pendulum mass, respectively, l is the pendulum length and ϕ is its angle from the vertical axis. Force u is the control input, x is the cart displacement and g is gravity. We use the Runge-Kutta 4th-order method to discretize (47) with a sampling time $dt = 0.1s$ and cast it as $\mathbf{s}^+ = \mathbf{f}(\mathbf{s}, \mathbf{a}) + \boldsymbol{\xi}$, where $\mathbf{s} = [x, \dot{x}, \phi, \dot{\phi}]^\top$ is the state, $\mathbf{a} = u$ is the input, $\boldsymbol{\xi}$ is a Gaussian noise and \mathbf{f} is a nonlinear function representing (47) in discrete time. We consider the state constraint $x \geq 0$, discount factor $\gamma = 0.95$ and the following MDP stage cost to stabilize the system at the origin while penalizing the system constraint:

$$\ell(\mathbf{s}, \mathbf{a}) = \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix}^\top \begin{bmatrix} I_4 & 0 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} \mathbf{s} \\ \mathbf{a} \end{bmatrix} + \lambda \max(-x, 0), \quad (48)$$

where λ is a large constant value introduced to model the state constraint as a soft constraint. In the MPC scheme, we use the linear model $\mathbf{s}^+ = \hat{A}\mathbf{s} + \hat{B}\mathbf{a}$ obtained by linearizing \mathbf{f} at the origin. We provide a parameterized quadratic stage and terminal cost and select prediction horizon $N = 20$. We use the deterministic policy gradient method to minimize the performance function $J(\boldsymbol{\theta}) := \mathbb{E}_{s_0}[\hat{V}_N^\theta(s_0)]$, and we run a simulation for 1000 learning steps of the policy gradient method. Figure 7 shows the value function over the learning steps for a fixed initial state. This illustrates that RL successfully manages to reduce J throughout the iterates, therefore tuning MPC as desired.

Figure 8 shows the states and input trajectories of the real system corresponding to the 1000th learning step of the policy gradient method. The MPC scheme with the positive definite stage cost and other stability conditions in the terminal cost, terminal constraint is able to deliver the stabilizing policy for the closed-loop system for the small enough model error [8]. Note that the terminal cost and constraint conditions can be relaxed for the large enough MPC horizon [25]. Figure 9 compares the state constraint violation for $x \geq 0$ in the first and the last (1000th) learning step. As one can see, RL reduces the state constraint violation. Note that, we have used a common MPC formulation as (27) in this example. However, one can use robust MPC to avoid constraint violation as shown in [22].

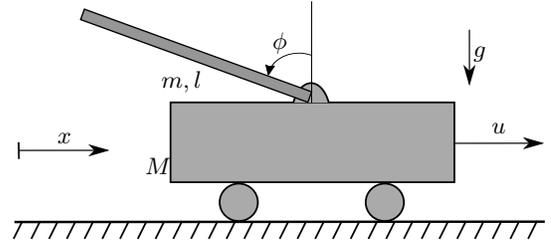


Fig. 6. The cart-pendulum system. We use $M = 0.5\text{kg}$, $m = 0.2\text{kg}$, $l = 0.3\text{m}$ and $g = 9.8\text{m/s}^2$ for the simulation.

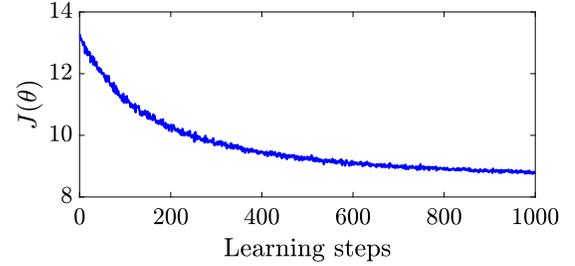


Fig. 7. The closed-loop performance of the MPC scheme over RL-steps.

D. Learning based MPC: Economic stage cost

In this example, we investigate an economic cost in the real system with bias optimality criterion. We use a parameterized MPC scheme with a parameterized storage function as a function approximator in the Q-learning algorithm. Continuously Stirred Tank Reactor (CSTR) is a common ideal reactor in chemical engineering, usually used for liquid-phase or multiphase reactions with fairly high reaction rates. The CSTR nonlinear dynamics can be written as follows (see [26]):

$$\begin{aligned} \dot{C}_A &= \frac{F}{V_R}(C_{A0} - C_A) - k_0 e^{-E/RT} C_A^2 \\ \dot{T} &= \frac{F}{V_R}(T_0 - T) - \frac{\Delta H k_0}{\rho_R C_p} e^{-E/RT} C_A^2 + \frac{q}{\rho_R C_p V_R}, \end{aligned} \quad (49)$$

where T denotes the temperature of the reactor contents, C_A is the concentration of A in the reactor, F is the flow rate, and q is the heat rate. The remaining notation definitions and process parameter values are given in e.g., [20]. Then $\mathbf{s} = [C_A, T]^\top$ and $\mathbf{a} = [F, q]^\top$ are the state and input of the system, respectively. The input \mathbf{a} must satisfy the following inequality:

$$[0, -2e5]^\top \leq \mathbf{a} \leq [10, 2e5]^\top \quad (50)$$

An economic stage cost is defined as follows:

$$\ell(\mathbf{s}, \mathbf{a}) = -\eta \underbrace{F(C_{A0} - C_A)}_{:=r} + \beta q \quad (51)$$

where η and β are positive constants, and r is the production rate. This cost maximizes the production rate and minimizes the energy consumption of the production (the second term). We consider $\eta = 1.7e4$ and $\beta = 1$ for the simulation. Sampling time 0.02h is used to discretize the system (49).

We use an MPC scheme with a neural network-based storage function and parameterized stage cost and terminal cost and we denote the adjustable parameters by $\boldsymbol{\theta}$. Then we use Q-learning in order to update the parameters $\boldsymbol{\theta}$. Figure 10 (left) illustrates the value function $\hat{V}_N^\theta(s_0)$. It can be seen that the parameterized value function is decreasing during the learning. Figure 10 (right) shows the convergence of the parameters.

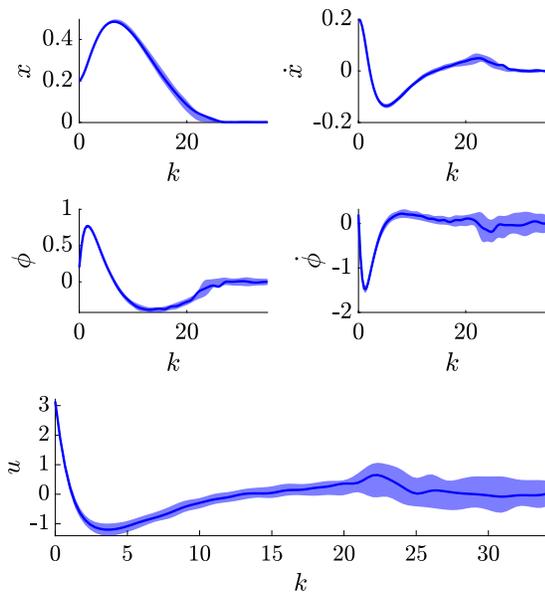


Fig. 8. States and input trajectories of the real system for the last learning step.

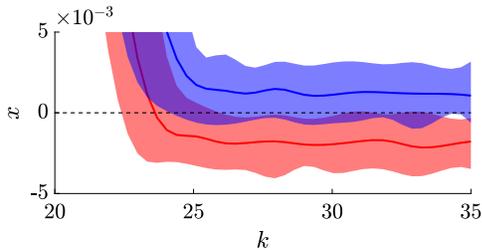


Fig. 9. Violation of the state constraint $x \geq 0$ in the first step (red) and the last step (blue).

VII. CONCLUSION

In this paper, we showed that a finite-horizon OCP can capture the optimal policy and value functions of any MDPs with either discounted or undiscounted cost even if we use an inexact model in the OCP. We showed that an MPC scheme can be interpreted as a particular case of the OCP where we use a deterministic model to avoid computational complexity. In practice, we proposed the use of a parameterized MPC scheme to provide a structured function approximator for the RL techniques. RL algorithms then can be used in order to tune the MPC parameters to achieve the best closed-loop performance. We verified the theorems in an LQR case and investigated some nonlinear examples to illustrate the efficiency of the method numerically.

REFERENCES

- [1] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [2] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, p. 387–395.
- [3] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.
- [4] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2007, vol. 703.
- [5] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [6] D. P. Bertsekas, *Approximate dynamic programming*. Citeseer, 2008.
- [7] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

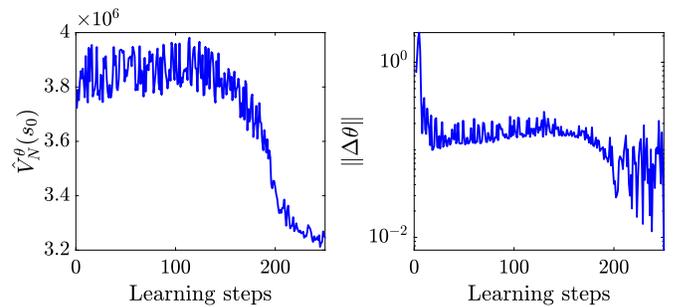


Fig. 10. (Left:) The MPC-based value function $\hat{V}_N^\theta(s_0)$ during the learning. (Right:) Convergence of the norm of the parameters during the Q-learning steps.

- [8] J. B. Rawlings, D. Q. Mayne, and M. Diehl, *Model predictive control: theory, computation, and design*. Nob Hill Publishing Madison, WI, 2017, vol. 2.
- [9] S. Gros and M. Zanon, "Data-driven economic NMPC using Reinforcement Learning," *IEEE Transactions on Automatic Control*, vol. 65, no. 2, pp. 636–648, 2019.
- [10] A. B. Kordabad, W. Cai, and S. Gros, "Multi-agent battery storage management using MPC-based reinforcement learning," in *2021 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2021, pp. 57–62.
- [11] A. B. Kordabad, W. Cai, and S. Gros, "MPC-based reinforcement learning for economic problems with application to battery storage," in *2021 European Control Conference (ECC)*. IEEE, 2021, pp. 2573–2578.
- [12] H. N. Esfahani, A. B. Kordabad, and S. Gros, "Reinforcement learning based on MPC/MHE for unmodeled and partially observable dynamics," in *2021 American Control Conference (ACC)*, 2021, pp. 2121–2126.
- [13] M. Granzotto, R. Postoyan, L. Buşoniu, D. Nešić, and J. Daafouz, "Finite-horizon discounted optimal control: stability and performance," *IEEE Transactions on Automatic Control*, vol. 66, no. 2, pp. 550–565, 2020.
- [14] M. Zanon and S. Gros, "A new dissipativity condition for asymptotic stability of discounted economic MPC," *Automatica*, vol. 141, p. 110287, 2022.
- [15] M. Zanon, S. Gros, and M. Palladino, "Stability-constrained Markov decision processes using MPC," *Automatica*, vol. 143, p. 110399, 2022.
- [16] S. Mahadevan, "Average reward reinforcement learning: Foundations, algorithms, and empirical results," *Machine learning*, vol. 22, no. 1, pp. 159–195, 1996.
- [17] S. Gros and M. Zanon, "Economic MPC of Markov decision processes: Dissipativity in undiscounted infinite-horizon optimal control," *Automatica*, vol. 146, p. 110602, 2022.
- [18] R. Postoyan, L. Buşoniu, D. Nešić, and J. Daafouz, "Stability analysis of discrete-time infinite-horizon optimal control with discounted cost," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2736–2749, 2016.
- [19] E. C. Kerrigan and J. M. Maciejowski, "Soft constraints and exact penalty functions in model predictive control," in *Control 2000 Conference, Cambridge*. Citeseer, 2000, pp. 2319–2327.
- [20] A. B. Kordabad and S. Gros, "Q-learning of the storage function in economic nonlinear model predictive control," *Engineering Applications of Artificial Intelligence*, vol. 116, p. 105343, 2022.
- [21] A. B. Kordabad, H. N. Esfahani, A. M. Lekkas, and S. Gros, "Reinforcement learning based on scenario-tree MPC for ASVs," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 1985–1990.
- [22] M. Zanon and S. Gros, "Safe reinforcement learning using robust mpc," *IEEE Transactions on Automatic Control*, 2020.
- [23] M. S. Santos and J. Vigo-Aguiar, "Analysis of a numerical dynamic programming algorithm applied to economic models," *Econometrica*, pp. 409–426, 1998.
- [24] L. Grüne, C. M. Kellett, and S. R. Weller, "On a discounted notion of strict dissipativity," *IFAC-PapersOnLine*, vol. 49, no. 18, pp. 247–252, 2016.
- [25] A. Jadbabaie and J. Hauser, "On the stability of receding horizon control with a general terminal cost," *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 674–678, 2005.
- [26] Xinchun Li, Liqin Zhang, M. Nakaya, and A. Takenaka, "Application of economic MPC to a CSTR process," in *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2016, pp. 685–690.