Equivalency of Optimality Criteria of Markov Decision Process and Model Predictive Control

Arash Bahari Kordabad, Mario Zanon, Sébastien 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 limit the computational complexity. In practice, Reinforcement Learning algorithms can then be used to tune the parameterized MPC scheme. We verify the proposed 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. 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. 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 yield 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. This issue is known as "curse of dimensionality" in the literature. 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) and approximate DP are two common techniques that tackle these difficulties.

RL is a powerful, model-free method. Indeed, RL offers 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. Deep Neural Networks (DNNs) are a common function approximator 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. DNNs usually need a large number of tunable parameters. Moreover, 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. 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, 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.,) and partially observable MDPs. This work extends the results of to the case of undiscounted MPC. More specifically, we show that an undiscounted MPC is able to capture the optimal policy of an MDP with either discounted or undiscounted criteria even if the MPC model is inaccurate.

Stability for discounted MPC schemes is challenging, and for a finite-horizon problem, it is shown 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 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.

The equivalency of MDPs criteria (discounted and undiscounted) has been recently discussed in 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. In this paper, we first 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 MDPs, 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) and detail discounted and undiscounted criteria of optimality. We consider an MDP with the following transition probability measure:

$$\rho (s^{+}, a),$$

where $s \in \mathcal{X} \subseteq \mathbb{R}^n$, $a \in \mathcal{U} \subseteq \mathbb{R}^m$, and $s^{+}$ are the current state, input, and subsequent state, respectively, and $\rho : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \rightarrow \mathbb{R}^\geq0$ is the probability measure associated with this transition. The input $a$ applied to the system for a given state $s$ is selected by a deterministic policy $\pi : \mathcal{X} \rightarrow \mathcal{U}$. We denote $s_0, s_1, ..., s_{t-1}$ the (possibly stochastic) trajectories of the system under policy $\pi$, i.e., $s_{t+1} \sim \rho (|s_t, \pi (s_t))$, starting from $s_0 = s$, $\forall \pi$. We further denote the measure associated with such trajectories as $\tau^\pi$.

A. Discounted MDPs

In the discounted setting, solving the MDP consists in finding an optimal policy $\pi^\star$, solution of the following discounted infinite-horizon OCP:

$$V^\star (s) := \min_{\pi} V^\pi (s) := E_{\tau^\pi} \left[ \sum_{k=0}^\infty \gamma^k \ell (s_k, \pi (s_k)) \right],$$

for all initial state $s_0 = s$, where $V^\star : \mathcal{X} \rightarrow \mathbb{R}$ is the optimal value function, $V^\pi$ is the value function of the Markov Chain in closed-loop with policy $\pi$, $\ell : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ is the stage cost function of the real system and $\gamma \in (0, 1]$ is the discount factor. The expectation $E_{\tau^\pi}$ is taken over the distribution underlying the Markov Chain $\{W_t\}$ in closed-loop with policy $\pi$.

The action-value function $Q^\pi (s, a)$ and advantage function $A^\pi (s, a)$ associated to (2) are defined as follows:

$$Q^\pi (s, a) := \ell (s, a) + \gamma E_{\rho} \left[ V^\pi (s^{+}) | s, a \right],$$

$$A^\pi (s, a) := Q^\pi (s, a) - V^\star (s).$$

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

$$V^\star (s) = Q^\pi (s, \pi^\star (s)), \text{ } \forall \pi \in \mathcal{X},$$

and

$$0 = \min_{\alpha} A^\pi (s, a), \text{ } \pi^\star (s) \in \arg \min_{\alpha} A^\pi (s, a), \text{ } \forall \pi \in \mathcal{X}.$$
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:

\[
\hat{\rho} [s^+ | s, a].
\] (8)

In order to distinguish it from the real system trajectory, let us denote \(\hat{s}_0, \hat{s}_1, \ldots\) the (possibly stochastic) trajectories of the state transition model \([8]\) under policy \(\pi\), i.e., \(\hat{s}_{k+1} \sim \hat{\rho} [\cdot | \hat{s}_k^\pi, \pi(\hat{s}_k^\pi)]\), starting from \(\hat{s}_0 = s, \forall \pi\). We further denote the measure associated with such trajectories as \(\hat{\pi}\).

It has been shown in \([21]\) 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 \([19]\). This observation is well-known in MPC of deterministic systems \([22]\). Therefore, in this paper, we are interested in using an undiscounted OCP for the model \([8]\) 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 the imperfect system model \([9]\) which we will formulate in \([9]\). 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 \([9]\):

\[
\hat{V}_N^\pi(s) = \min_{\pi} \hat{V}_N^\pi(s) := \mathbb{E}_{\pi} \left[ \hat{T}(\hat{s}_N) \right] + \sum_{k=0}^{N-1} \hat{L} (\hat{s}_k^\pi, \pi(\hat{s}_k^\pi)) \right],
\] (9)

with initial state \(\hat{s}_0 = s\), where \(N\) is the horizon length, \(\hat{T}\), \(\hat{L}\), \(\hat{V}_N^\pi\) and \(\hat{V}_N^\pi\) are the terminal cost, the stage cost, the optimal value function and the value function of the policy \(\pi\) associated to model \([9]\), respectively. The expectation \(\mathbb{E}_{\pi}\) in \([9]\) is taken over undiscounted closed-loop Markov Chain \([9]\) with policy \(\pi\). We denote \(\pi^*_N\) the optimal policy resulting from \([9]\). Moreover, the action-value function \(Q^*_N\) associated to \([9]\) is defined as follows:

\[
Q^*_N(s, a) := \hat{L}(s, a) + \mathbb{E}_{\hat{\pi}} \left[ \hat{V}^\pi_{N-1}(s^+) | s, a \right].
\] (10)

The next assumption expresses a requirement on the model trajectories \(\hat{s}_0, \hat{s}_1, \ldots\) with the optimal policy \(\pi^*\) which allows us to develop the theoretical results of this paper.

Assumption 1. The set

\[
\mathcal{S} = \{ s \in \mathcal{X} | \mathbb{E} [V^\pi(\hat{s}_k^\pi)] < \infty, \forall k \leq \hat{N} \}\]

is non-empty for a given \(\hat{N} \in \mathbb{I}\).

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. This assumption plays a vital role in the derivation of our main result.

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 \(\hat{N} \geq N\). Then, there exist a terminal cost \(\hat{T}\) and a stage cost \(\hat{L}\) such that the following identities hold, \(\forall \gamma\):

(i) \(\hat{V}_N^\pi(s) = V^*(s), \forall N \in \mathbb{I}_{\geq 0}, \forall s \in \mathcal{S}\)

(ii) \(\pi_N^*(s) = \pi^*(s), \forall N \in \mathbb{I}_{\geq 1}, \forall s \in \mathcal{S}\)

(iii) \(\hat{Q}_N^*(s, a) = Q^*(s, a), \forall N \in \mathbb{I}_{\geq 1}, \forall s \in \mathcal{S}, \forall a \in \mathcal{U}\) such that \(|\mathbb{E}_{\hat{\pi}}[V^*(s^+)]| < \infty\)

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

\[
\hat{T}(s) = V^*(s) \quad (12a)
\]

\[
\hat{L}(s, a) = \begin{cases} 
Q^*(s, a) - \mathbb{E}_{\hat{\pi}}[V^*(s^+)] | s, a & \text{if } |\mathbb{E}_{\hat{\pi}}[V^*(s^+)]| < \infty \\
\infty & \text{otherwise}
\end{cases} \quad (12b)
\]

Then \([12]\) is a direct consequence of Assumption \([1]\) and \([12a]\) for \(N = 0\). For \(N > 0\), under Assumption \([1]\) the terminal and stage costs in \([9]\) have a finite expected value for all \(\hat{s}_0 \in \mathcal{S}\). By substitution of \([12]\) in \([9]\) and using telescopic sum, we have:

\[
\hat{V}_N^\pi(s) = \mathbb{E} \left[ \hat{T}(\hat{s}_N^\pi) + \sum_{k=0}^{N-1} \hat{L}(\hat{s}_k^\pi, \pi(\hat{s}_k^\pi)) \right], \quad \text{with initial state } \hat{s}_0 = s, \quad \hat{V}_N^\pi\]
Assumption 2.

which corresponds to the cost modification discussed in [19].

Moreover, from (10) and (12b), for any inputs $\alpha \in \mathcal{U}$ such that $|E_{\rho}[V^*(s^+)|s, \alpha]| < \infty$, we have:

$$
\hat{Q}^*(s, \alpha) = \hat{L}(s, \alpha) + E_{\rho} \left[ V^* - V^* \left| s^+, s, \alpha \right. \right] 
$$

where the last inequality is obtained by noting that (i) also implies $V^* - V^*$. This directly yields (11).

Theorem 1 states that, independent of the discount factor $\gamma$, it is possible to find a finite-horizon OCP that provides the optimal policy and optimal value functions of a discounted MDP if an inexact model is used. We observe that the setup of this paper has been analyzed in [19], under the assumption of a perfect model, i.e., $\rho|s, \alpha| = \rho|s, \alpha|$. In that case [12b] reads:

$$
\hat{L}(s, \alpha) = \ell(s, \alpha) + (\gamma - 1)E_{\rho}[V^*|s, \alpha], \forall s \in S, (18)
$$

which corresponds to the cost modification discussed in [19].

B. Infinite-horizon OCP

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

Assumption 2. We assume that the optimal value function converges to a constant and finite value with model (5) under the optimal policy $\pi^*$. I.e.:

$$
-\infty < \lim_{N \to \infty} E_{\rho} \left[ V^*(\hat{s}_N^+) \right] = \hat{v}_\infty < \infty
$$

Assumption 2 can be interpreted as some forms of the stability condition on the model dynamics under the optimal policy $\pi^*$. We will further discuss this assumption in Section IV.

In this section, we consider the following undiscounted value function without terminal cost:

$$
\hat{V}^*_\infty(s) := \min_{\pi} \hat{V}^*_\infty(s) := \lim_{N \to \infty} E_{\pi^*} \left[ \sum_{k=0}^{N-1} \hat{L}(\hat{s}_k^+, \pi(\hat{s}_k^+)) \right]
$$

with initial state $\hat{s}_0^\pi = s$. We denote the optimal policy solution of (20) as $\pi^*_\infty(s)$. We then define the optimal action-value function $Q^*_\infty$ associated to (20) as follows:

$$
Q^*_\infty(s, \alpha) = \hat{L}(s, \alpha) + E_{\rho} \left[ V^*_\infty(s^+)\right|s, \alpha
$$

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 $\forall s \in S, \forall \gamma$:

(i) $\pi^*_\infty(s) = \pi^*(s)$

(ii) $\hat{V}^*_\infty(s) = V^*(s) - \hat{v}_\infty$

(iii) $Q^*_\infty(s, \alpha) = Q^*(s, \alpha) - \hat{v}_\infty$, for the inputs $\alpha \in \mathcal{U}$ such that $|E_{\rho}[V^*(s^+)|s, \alpha]| < \infty$

if the stage cost $\hat{L}$ is selected according Equation (12b).

Proof. Using stage cost $\hat{L}$ in (12b), we have:

$$
\hat{V}^*_\infty(s) = \lim_{N \to \infty} E_{\pi^*} \left[ \sum_{k=0}^{N-1} Q^*(s_k^+, \pi(s_k^+)) \right]
$$

which completes the proof.
Theorem 2 extends Theorem 1 to the case of an infinite horizon with zero terminal cost and states that under the additional Assumption 2. This assumption 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 (9) 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 (23), 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 for each initial state. In practice, this amounts to deploying the finite-horizon OCP in an MPC framework, i.e., in 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 \( \theta \) such that RL methods can be deployed to tune \( \theta \) 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 \( \hat{\rho} [s^+|s,a] \) is a deterministic model, i.e.:

\[
\hat{\rho} [s^+|s,a] = \delta (s^+ - f_\theta(s,a))
\]  

where \( \delta(\cdot) \) is the Dirac measure and \( f_\theta(s,a) \) is a parameterized deterministic (possibly nonlinear) model.

We approximate the modified costs \( \hat{L} \) and \( \hat{T} \) by parametric functions \( L_\theta \) and \( T_\theta \), 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 [23]. We therefore formulate the MPC finite-horizon OCP as:

\[
\hat{V}_N^\theta(s) = \min_{\hat{a},\hat{s},\sigma} -\lambda_\theta(\hat{s}_0) + T_\theta(\hat{s}_N) + \mu_\ell^T \sigma_N + \sum_{k=0}^{N-1} L_\theta(\hat{s}_k, \hat{a}_k) + \mu_\ell^T \sigma_k \quad (27a)
\]

such that \( \hat{s}_{k+1} = f_\theta(\hat{s}_k, \hat{a}_k), \hat{s}_0 = s \), \( \hat{a}_k \in \mathcal{U}, 0 \leq \sigma_k, 0 \leq \sigma_N \), \( h_\theta(\hat{s}_k, \hat{a}_k) \leq \sigma_k^\ell, h_\theta(\hat{s}_N) \leq \sigma_N^\ast \), \( (27b) \) (27c) (27d)

where \( \hat{V}_N^\theta \) is the MPC-based parameterized value function, \( h_\theta(s,a) \) is a mixed input-state constraint, \( h_\theta^0(s) \) is the terminal constraint, \( \sigma_k \) and \( \sigma_N \) are slack variables guaranteeing the feasibility of the MPC scheme and \( \mu_\ell \) and \( \mu_\ell^T \) are constant vectors that ought to be selected sufficiently large [23]. 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 \( \lambda_\theta \) 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 \( -\lambda_\theta(\hat{s}_0) \) only depends on the current state, it does not modify the optimal policy. For more details, we refer the interested readers to [24, 12].

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 \( f_\theta \) introduces extra degrees of freedom to bring \( \hat{L} \) and \( \hat{T} \) closer to the functions that can be represented by \( L_\theta \) and \( T_\theta \). 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:

\[
\hat{\pi}_N^\theta(s) = \hat{a}_0^\ast(\theta, s), \quad (28)
\]

where \( \hat{a}_0^\ast \) is the solution of (27), corresponding to the first input \( a_0 \). Moreover, the parameterized action-value function \( \hat{Q}_\theta(s,a) \) based on MPC scheme (27) can be formulated as follows:

\[
\hat{Q}_N^\theta(s,a) := \min_{a,s,\sigma} \quad (27a)
\]

\[
\text{s.t.} \quad (27b) - (27d) \quad (29b)
\]

\[
a_0 = a. \quad (29c)
\]

Then one obtains the following identities:

\[
\hat{V}_N^\theta(s) = \min_a \hat{Q}_N^\theta(s, a), \quad (30a)
\]

\[
\hat{\pi}_N^\theta(s) = \arg \min_a \hat{Q}_N^\theta(s, a). \quad (30b)
\]
We can use RL techniques, such as Q-learning and policy gradient method to tune the parameters $\theta$ of parameterized MPC scheme [27] and approach the optimal parameter $\theta^*$. The use of RL for the tuning the MPC scheme can be found e.g., in [12], [25], [26].

The next section provides an analytical case study to illustrate the theoretical developments of this paper.

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 state system transition and stage cost are given as follows:

$$s^+ = As + Ba + e, \quad e \sim N(0, \Sigma),$$

(31a)

$$\ell(s, a) = \begin{bmatrix} s^T \\ a^T \end{bmatrix} \begin{bmatrix} T & N \\ N^T & R \end{bmatrix} \begin{bmatrix} s \\ a \end{bmatrix},$$

(31b)

with discount factor $\gamma$. One can verify the following optimal value functions:

$$V^*(s) = s^T Ss + \bar{v}_\infty,$$

(32)

$$Q^*(s, a) = \bar{v}_\infty + \begin{bmatrix} s^T \\ a^T \end{bmatrix} \begin{bmatrix} T + \gamma A^T SA & N + \gamma A^T SB \\ N^T + \gamma B^T SA & R + \gamma B^T SB \end{bmatrix} \begin{bmatrix} s \\ a \end{bmatrix},$$

(33)

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

$$T + \gamma A^T SA = S + (N + \gamma A^T SB)(K^*_\gamma)^T,$$

(33a)

$$(R + \gamma B^T SB)K^*_\gamma = N^T + \gamma B^T SA.$$

(33b)

Then $\pi^*(s) = -K^*_\gamma s$ and $\pi^*_\infty(s) = \pi^*(s) = -K^*_\gamma s$, where $K^*_\gamma = \text{lim}_{\gamma \to 1} K^*_\gamma$. We then consider a linear deterministic model:

$$s^+ = As + Ba,$$

(34)

and an undiscounted OCP with the following stage cost, defined accordingly to Equation (12b) as:

$$\hat{L}(s, a) = Q^*(s, a) - V^*(\hat{s}^+)$$

(35)

The Riccati equations for the undiscounted problem with the model (34) read as:

$$\hat{T} + \hat{A}^T \hat{S}A = \hat{S} + (\hat{N} + \hat{A}^T \hat{S}B)(K^*_\gamma)^T,$$

(36a)

$$(\hat{R} + \hat{B}^T \hat{S}B)K^*_\gamma = \hat{N}^T + \hat{B}^T \hat{S}A.$$

(36b)

with the optimal policy $\pi^*_\infty(s) = -K^*_\gamma s$ and the optimal value function $V^*_\infty(s) = s^T \hat{S}s$. From (35), we have:

$$T + \gamma A^T SA - \hat{A}^T \hat{S}A = \hat{T},$$

(37a)

$$N + \gamma A^T SB - \hat{A}^T \hat{S}B = \hat{N},$$

(37b)

$$R + \gamma B^T SB - \hat{B}^T \hat{S}B = \hat{R}.$$
Fig. 1. Optimal value functions resulting from the discounted real system and undiscounted MPC scheme with the wrong model.

Fig. 2. Optimal policy functions resulting from the discounted real system and undiscounted MPC scheme with the wrong model.

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 \( \xi \sim U[-0.5, 0.5] \) has a uniform distribution and \( s_k := [s_k(1), s_k(2)]^T \) 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 and Figure 4 show the optimal value 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:

\[
\dot{s}_{k+1} = f_\theta(\hat{s}_k, \hat{a}_k) = \hat{s}_k + \begin{bmatrix} \frac{\partial}{\partial \hat{s}_k(1)} \hat{s}_k(2) \\ \frac{\partial}{\partial \hat{a}_k(1)} \end{bmatrix} \delta t + \begin{bmatrix} 0 \\ \delta t \frac{m l^2}{m \ell} \end{bmatrix} \hat{a}_k \tag{44}
\]

where \( \hat{s}_k := [\hat{s}_k(1), \hat{s}_k(2)]^T \) and \( \hat{a}_k \) are the model state and input. Moreover, we consider an uncertain \( l \) with a adjustable parameter \( \theta_l \), with 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:

\[
T_\theta(s) = s^\top G s, \quad L_\theta(s, a) = \begin{bmatrix} s \\ a \end{bmatrix}^\top H \begin{bmatrix} s \\ a \end{bmatrix} \tag{45}
\]

where \( G \) and \( H \) are parametric positive definite matrices. Then the parameters vector \( \theta \) gathers all of the adjustable parameters as:

\[
\theta = \{\theta_l, G, H\} \tag{46}
\]

Figure 5 shows the difference between the MPC value function \( \hat{V}_N^\theta \) and the optimal value function \( V^\star \) computed by DP. The blue and red curves represent this difference at the beginning of the learning and after 500 learning steps, respectively. Figure 6 shows the difference between the MPC policy \( \hat{\pi}_N^\theta \) and the optimal policy \( \pi^\star \) computed by DP. 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 \( V^\star \)

Fig. 3. Optimal Value function resulting from ADP.

Fig. 4. Optimal Policy function resulting from DP.
and $\pi^*$, as per Theorem 1, do not have a quadratic form, hence the selected MPC formulation cannot capture them exactly. The green curves in Figures 5 and 6 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 7 illustrates the closed-loop performance of the system under the MPC policy $\hat{\pi}^*_\theta$. 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 8 in order to illustrate the proposed method in a constrained tracking problem. The dynamics are given by

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

$$
\frac{1}{3}ml^2\dddot{\phi} + \frac{1}{2}ml\dot{\phi} \cos \phi = -\frac{1}{2}mgl \sin \phi,
$$

where $M$ and $m$ are the cart mass and pendulum mass, respectively, $l$ is the pendulum length and $\phi$ is its angle from the vertical axis. Force $u$ is the control input, $x$ is the cart displacement and $g$ is gravity. We use Runge-Kutta 4-th-order method to discretize (47) with a sampling time $dt = 0.1s$ and cast it in the form of $s^{+} = f(s, a) + \xi$, where $s = [x, \dot{x}, \phi, \dot{\phi}]^\top$ is the state, $a = u$ is the input, $\xi$ is a Gaussian noise and $f$ 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:

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

where $\lambda$ is a large constant value introduced to model the state constraint as a soft constraint.

In the MPC scheme, we use the linear model $s^{+} = \dot{A}s + Ba$ obtained by linearizing $f$ at the origin. We provide a parametrized quadratic stage and terminal cost and select prediction horizon $N = 20$.

We use the deterministic policy gradient method to minimize the performance function $J(\theta) := \mathbb{E}_{s_0}[\hat{V}^*_N(s_0)]$, and we run a simulation for 1000 learning steps. Figure 9 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 10 shows the states and input trajectories of the real system corresponding the 1000-th learning step. The MPC
Fig. 9. The closed-loop performance of the MPC scheme over RL-steps.

Fig. 10. States and input trajectories of the real system for the last learning scheme with the positive definite stage cost and other stability conditions in the terminal cost and terminal constraint is able to deliver the stabilizing policy for the closed-loop system for the small enough model error [11]. Note that the terminal cost and constraint conditions can be relaxed for the large enough MPC horizon [29].

Figure 11 compares the state constraint violation for \( x \geq 0 \) in the first and last (1000th) learning step. As one can see, RL reduces the state constraint violation.

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 [30]):

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

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 Table I. Then \( s = [C_A, T]^\top \) and \( a = [F, q]^\top \) are the state and input of the system, respectively. The input \( a \) must satisfy the following inequality:

\[
[0, -2e5]^\top \leq a \leq [10, 2e5]^\top
\]

An economic stage cost is defined as follows:

\[
\ell(s, a) = -\alpha F(C_{A0} - C_A) + \beta q
\]

where \( \alpha \) and \( \beta \) 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 \( \alpha = 1.7e4 \) 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 networks based storage function and parameterized stage cost and terminal cost and we denote the adjustable parameters by \( \theta \). Then we use Q-learning in order to update the parameters \( \theta \). Figure 12 shows the convergence of the parameters vector during learning. Figure 13 illustrates the value function \( \hat{V}_N^\theta(s_0) \). It can be seen that the parameterized value function is decreasing during the learning.

| Symbol | Description | Value |
|--------|-------------|-------|
| \( C_{A0} \) | Feed concentration of \( A \) \( 3.5 \text{kmol/m}^3 \) | |
| \( T_0 \) | Feedstock temperature \( 300 \text{K} \) | |
| \( V_R \) | Reactor fluid volume \( 1.0 \text{m}^3 \) | |
| \( E \) | Activation energy \( 5.0e4 \text{kJ/kmol} \) | |
| \( k_0 \) | Pre-exponential rate factor \( 8.46e6 \text{m}^3/\text{kmolh} \) | |
| \( \Delta H \) | Reaction enthalpy change \( -1.16e4 \text{kJ/kmol} \) | |
| \( C_p \) | Heat capacity \( 0.231 \text{kJ/kgK} \) | |
| \( \rho_R \) | Density \( 1000 \text{kg/m}^3 \) | |
| \( R \) | Gas constant \( 8.314 \text{kJ/kmolK} \) | |

The CSTR nonlinear dynamics can be written as follows (see [20]):

\[
\begin{align*}
\dot{C}_A &= \frac{F}{V_R}(C_{A0} - C_A) - k_0 e^{-E/RT} C_A^2 \\
\dot{T} &= \frac{F}{V_R}(T_0 - T) - \frac{\Delta H k_0}{\rho_R C_p} e^{-E/RT} C_A^2 + \frac{q}{\rho_R C_p V_R},
\end{align*}
\]
We verified the theorems in an LQR case and investigated some with either discounted or undiscounted cost even if we use an (SARLEM).

In practice, we proposed the used of a parameterized MPC scheme to provide a structured function approximator for the RLcapture the optimal policy and value functions of any MDPs.

\[ \hat{V}_N(\theta) \]

Learning steps

\[ 10^6 \]

\[ 0 \]

\[ 50 \]

\[ 100 \]

\[ 150 \]

\[ 200 \]

\[ 250 \]

\[ \Delta \theta \]

Learning steps

\[ 4 \times 10^6 \]

\[ 0 \]

\[ 50 \]

\[ 100 \]

\[ 150 \]

\[ 200 \]

\[ 250 \]

\[ V^*_N(\theta_0) \]

Fig. 12. Convergence of the norm of the parameters vector during the learning.

Fig. 13. The MPC-based value function $V^*_N(\theta_0)$ during the learning.

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 used 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.

Acknowledgment

This work was funded by the Research Council of Norway (RCN) project “Safe Reinforcement Learning using MPC” (SARLEM).

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 International Conference on Machine Learning. JMLR.org, 2014, p. 1-387–1-395.

[3] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in Advances in neural information processing systems, 2000, pp. 1057–1063.

[4] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, Dynamic programming and optimal control. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.

[5] W. B. Powell, Approximate Dynamic Programming: Solving the curses of dimensionality. John Wiley & Sons, 2007, vol. 703.

[6] J. Rust, “Using randomization to break the curse of dimensionality,” Econometrica: Journal of the Econometric Society, pp. 487–516, 1997.

[7] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.

[8] D. P. Bertsekas, Reinforcement learning and optimal control. Athena Scientific Belmont, MA, 2019.

[9] D. P. Bertsekas, Approximate dynamic programming. Citeeseer, 2008.

[10] K. Arulkumaran, M. P. Deisenroth, M. Brandage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 26–38, 2017.

[11] J. B. Rawlings, D. Q. Mayne, and M. Diehl, Model predictive control: theory, computation, and design. Nob Hill Publishing Madison, WI, 2017, vol. 2.

[12] 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.

[13] A. B. Kordabad, H. N. Esfahani, A. M. Lekkas, and S. Gros, “Reinforcement learning based on scenario-tree mpc for avss,” in 2021 American Control Conference (ACC), 2021, pp. 1985–1990.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] M. Zanon and S. Gros, “A new dissipativity condition for asymptotic stability of discounted economic mpc,” Automatica, vol. 141, p. 110287, 2022.

[19] M. Zanon, S. Gros, and M. Palladino, “Stability-constrained markov decision processes using mpc,” Automatica, vol. 143, p. 110399, 2022.

[20] S. Mahadevan, “Average reward reinforcement learning: Foundations, algorithms, and empirical results,” Machine learning, vol. 22, no. 1, pp. 159–195, 1996.

[21] S. Gros and M. Zanon, “A dissipativity theory for undiscounted markov decision processes,” arXiv preprint arXiv:2104.10997, 2021.

[22] 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.

[23] E. C. Kerrigan and J. M. Maciejowski, “Soft constraints and exact penalty functions in model predictive control,” in Control 2000 Conference, Cambridge. Citeeseer, 2000, pp. 2319–2327.

[24] A. B. Kordabad and S. Gros, “Verification of dissipativity and evaluation of storage function in economic nonlinear mpc using q-learning,” IFAC-PapersOnLine, vol. 54, no. 6, pp. 308–313, 2021, 7th IFAC Conference on Nonlinear Model Predictive Control NMPC 2021.

[25] A. B. Kordabad, H. N. Esfahani, A. M. Lekkas, and S. Gros, “Reinforcement learning based on scenario-tree mpc for avss,” in 2021 American Control Conference (ACC). IEEE, 2021, pp. 1985–1990.

[26] 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.

[27] M. S. Santos and J. Vigo-Aguiar, “Analysis of a numerical dynamic programming algorithm applied to economic models,” Econometrica, pp. 409–426, 1998.

[28] L. Grün, C. M. Kellett, and S. R. Weller, “On a discounted notion of strict dissipativity,” IFAC-PapersOnLine, vol. 49, no. 18, pp. 247–252, 2016.
[29] 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.

[30] 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.