2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Policy iteration for discrete-time systems with discounted costs: stability and near-optimality guarantees

Jonathan de Brusse[1], Mathieu Granzotto[2], Romain Postoyan[1] and Dragan Nešić[2]

***Abstract*— Given a discounted cost, we study deterministic discrete-time systems whose inputs are generated by policy iteration (PI). We provide novel near-optimality and stability properties, while allowing for non-stabilizing initial policies. That is, we first give novel bounds on the mismatch between the value function generated by PI and the optimal value function, which are less conservative in general than those encountered in the dynamic programming literature for the considered class of systems. Then, we show that the systems in closed-loop with policies generated by PI are stabilizing under mild conditions, after a finite (and known) number of iterations.**

I. Introduction

Dynamic programming provides powerful methods to generate near-optimal inputs for general dynamical systems and cost functions [2]. To make the best out of dynamic programming algorithms in a control engineering context, it is often essential to endow the obtained closed-loop system with stability guarantees. Various results exist in the literature ensuring stability properties for systems controlled by dynamic programming, both in continuous-time and discrete-time, see, e.g., [6,10,11,14,17,19]. The vast majority of these works focus on *undiscounted* cost functions. However, discounted costs are ubiquitous in dynamic programming and reinforcement learning [2,18], because of the favorable properties the discount endows to Bellman operators, like contractivity, see, e.g., [2]. We may also consider discounted costs because the problem at hand calls for it (e.g., economic inflation); or when a policy leading to a finite cost is known only in the discounted case. It is therefore important to provide stability guarantees for systems controlled by dynamic programming algorithms with discounted costs.

In [6,7] stability results are provided for discrete-time systems controlled by value iteration with discounted costs. Results for discounted policy iteration (PI) are only available for linear systems with quadratic costs (LQ) [13] where the discount factor is not fixed but increases with the number of iterations, as far as we know. In this work, we consider general deterministic discrete-time systems and costs with fixed discount factors. Our main goal is to establish stability properties when the inputs are generated by PI as well as novel near-optimality bounds.

We make several assumptions for this purpose. We first assume that an optimal sequence of inputs exists for any initial state and is stabilizing, which is very natural in the context of this work. We also assume that PI is recursively feasible in the sense that the optimization problem solved at each iteration is guaranteed to always admit a solution as customary in the literature [15]; if this is not the case we can resort to the modification of PI advocated in [8] and our results apply *mutatis mutandis*. On the other hand, the initial policy for PI is required to give a bounded finite cost, which does not mean that it is necessarily stabilizing because of the discount factor. This allows to relax the conditions of the related literature, see, e.g. [8,10,14], which require the initial policy to be stabilizing. This is one additional possible reason to consider discounted costs, i.e., to remove the need for an initial stabilizing policy for PI. Finally, the system needs to satisfy a detectability property with respect to the stage cost, which is also very natural as we aim to establish stability properties.

We use a generic measuring function to define stability as in e.g., [9,17], which is useful to study the stability of the origin and of more general attractors in a unified way. By exploiting the assumed stability property verified by the system in closed-loop with optimal controllers, novel near-optimality properties are deduced. Indeed, less conservative bounds on the mismatch between the value function generated by PI and the optimal value function compared to [2,16] are obtained by exploiting the properties of the considered class of systems. Afterwards, we establish via a Lyapunov analysis that PI generates stabilizing policies, provided a sufficient number of iterations. This is so even in the absence of stability properties of the initial policy as already mentioned. In particular, we show that the closed-loop system controlled by PI enjoys a semiglobal practical stability property where the adjustable parameter is the number of iterations. We provide an explicit relationship between the number of iterations required, the set of initial conditions and the guaranteed ultimate bound. By strengthening the assumptions, a global exponential stability property is derived and easy-to-compute lower bound on the discount factor and the number of iteration are given. We illustrate our results by means of two examples: the LQ problem and a nonholonomic integrator.

The rest of the paper is organized as follows. Prelim-

* This work was funded by Lorraine Université d'Excellence LUE, supported by the ANR under grant OLYMPIA ANR-23-CE48-0006 and the Australian Research Council under the Discovery Project DP210102600.

[1] J. de Brusse and R. Postoyan are with the Université de Lorraine, CNRS, CRAN, F-54000, Nancy, France. (emails: jonathan.de-brusse@univ-lorraine.fr, romain.postoyan@univ-lorraine.fr).

[2] D. Nešić and M. Granzotto are with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, VIC 3010, Australia (emails: dnesic@unimelb.edu.au, mathieu.granzotto@unimelb.edu.au).

inaries are recalled in Section II. The problem is formally stated in Section III. The standing assumptions are presented in Section IV. The main results are given in Section V. Examples are presented in Section VI before concluding in Section VII. Long proofs are omitted for space reasons and can be found in the extended version of this work [4].

II. Preliminaries

A. Notation

Let $\mathbb{R}$ be the set of real numbers, $\mathbb{R}_{\geq 0} := [0, +\infty)$, $\mathbb{Z}_{\geq 0} := \{0, 1, 2, ...\}$ and $\mathbb{Z}_{>0} := \{1, 2, ...\}$. We consider $\mathcal{K}$, $\mathcal{K}_\infty$ and $\mathcal{KL}$ functions as defined in [5, Section 3.5]. The identity map from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ is denoted by $\mathbb{I}$. Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, we use $f^{(k)}$ for the composition of function f to itself k times, where $k \in \mathbb{Z}_{\geq 0}$ and $f^{(0)} := \mathbb{I}$. We use $\lceil \cdot \rceil$ to denote the ceil function. The Euclidean norm of a vector $x \in \mathbb{R}^n$ with $n \in \mathbb{Z}_{>0}$ is denoted by $|x|$ and the distance of x to a non-empty set $\mathcal{A} \subseteq \mathbb{R}^n$ by $|x|_{\mathcal{A}} := \inf\{|x - y| : y \in \mathcal{A}\}$. For any $M \in \mathbb{R}^{n \times m}$ with $n, m \in \mathbb{Z}_{>0}$, $\|M\|$ is the spectral norm of the matrix M, i.e., $\|M\| = \sqrt{\rho(M^\top M)}$, where $\rho(M^\top M)$ is the spectral radius of the matrix $M^\top M$. If M is a symmetric matrix let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote its minimal and maximal eigenvalues, respectively. Given a set-valued map $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, a selection of S is a single-valued mapping $s : \operatorname{dom} S \to \mathbb{R}^m$ such that $s(x) \in S(x)$ for any $x \in \operatorname{dom} S$, we write $s \in S$ to denote a selection s of S for the sake of convenience. Finally, for an infinite sequence $\boldsymbol{u} = (u(0), u(1), ...)$ where $u(0), u(1), ... \in \mathbb{R}^m$ with $m \in \mathbb{Z}_{>0}$, $\boldsymbol{u}|_k$ stands for the truncation of $\boldsymbol{u}$ to its first $k \in \mathbb{Z}_{>0}$ steps, i.e., $\boldsymbol{u}|_k = \big(u(0), ..., u(k-1)\big)$ and we use the convention $\boldsymbol{u}|_0 = \emptyset$.

B. Plant Model and Cost Function

We consider nonlinear deterministic discrete-time systems given by

$$x(k+1) = f(x(k), u(k)), \qquad \forall k \in \mathbb{Z}_{\geq 0}, \tag{1}$$

where $x(k) \in \mathbb{R}^{n_x}$ is the state, $u(k) \in \mathcal{U}(x(k)) \subseteq \mathbb{R}^{n_u}$ is the control input at time step $k \in \mathbb{Z}_{\geq 0}$, $\mathcal{U}(x)$ is the non-empty set of *admissible* inputs at state $x \in \mathbb{R}^{n_x}$, and $n_x, n_u \in \mathbb{Z}_{>0}$. Ideally, we wish to find, for any given $x \in \mathbb{R}^{n_x}$, an infinite-length sequence of admissible inputs $\boldsymbol{u} = \big(u(0), u(1), ...\big)$ that minimizes the discounted infinite-horizon cost

$$J_\gamma(x, \boldsymbol{u}) := \sum_{k=0}^{\infty} \gamma^k \ell\big(\phi(k, x, \boldsymbol{u}|_k), u(k)\big), \tag{2}$$

where $\gamma \in (0, 1)$ is a discount factor, $\ell : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}_{\geq 0}$ is a non-negative stage cost and $\phi(k, x, \boldsymbol{u}|_k)$ is the solution to (1) at time $k \in \mathbb{Z}_{\geq 0}$, initialized at $x(0) = x \in \mathbb{R}^{n_x}$, with inputs given by $\boldsymbol{u}|_k$ and we use the convention $\phi(0, x, \boldsymbol{u}|_0) = x$. We assume that for any $x \in \mathbb{R}^{n_x}$, there exists a sequence of admissible inputs minimizing $J_\gamma(x, \cdot)$, i.e.,

$$V_\gamma^\star(x) := \min_{\boldsymbol{u}} J_\gamma(x, \boldsymbol{u}) < +\infty, \qquad \forall x \in \mathbb{R}^{n_x}, \tag{3}$$

as formalized in Section III-A. As a consequence, Bellman equation becomes

$$V_\gamma^\star(x) = \min_{u \in \mathcal{U}(x)} \big\{\ell(x, u) + \gamma V_\gamma^\star\big(f(x, u)\big)\big\} \qquad \forall x \in \mathbb{R}^{n_x}. \tag{4}$$

We can therefore define the non-empty set of optimal inputs for any state $x \in \mathbb{R}^{n_x}$ as

$$H_\gamma^\star(x) := \operatorname*{argmin}_{u \in \mathcal{U}(x)} \{\ell(x, u) + \gamma V_\gamma^\star\big(f(x, u)\big)\}. \tag{5}$$

Given (5), the closed-loop system (1) with optimal controller is given by

$$x(k+1) \in f\big(x(k), H_\gamma^\star(x(k))\big) =: F_\gamma^\star\big(x(k)\big) \quad \forall k \in \mathbb{Z}_{\geq 0}. \tag{6}$$

As (5) is a set-valued map, there may be non-unique optimal inputs at some state and, as a consequence, system (6) is a difference inclusion in general. For the sake of convenience, solutions to system (6) at time $k \in \mathbb{Z}_{\geq 0}$ are denoted as $\phi_\gamma^\star(k, x)$ when initialized at $x \in \mathbb{R}^{n_x}$.

Computing $H_\gamma^\star$ in (5) for the general dynamics in (1) and cost function (2) is notoriously hard. Dynamic programming provides algorithms to iteratively obtain feedback law, whose cost converges to the optimal one. We focus on PI in this work, which we recall in the next section. Before that, we introduce some notation. Given a policy $h : \mathbb{R}^{n_x} \to \mathbb{R}^{n_u}$ that is admissible, i.e., $h \in \mathcal{U}$, we denote the solution to system (1) in closed-loop with feedback law h at time $k \in \mathbb{Z}_{\geq 0}$ with initial condition $x(0) = x$ as $\phi(k, x, h)$. Likewise $J_\gamma(x, h)$ is the cost induced by h at initial state x, i.e., $J_\gamma(x, h) := \sum_{k=0}^{\infty} \gamma^k \ell\big(\phi(k, x, h), h\big(\phi(k, x, h)\big)\big)$.

III. Problem Statement

We recall PI in this section and we state the objectives of this work.

A. Policy Iteration

PI is given in Algorithm 1. Given $\gamma \in (0, 1)$ and an initial admissible policy h^0, PI generates at each iteration $i \in \mathbb{Z}_{\geq 0}$ a policy h_γ^{i+1} via the so-called improvement step in (PI.2). Policy h_γ^{i+1} is an arbitrary selection of H_γ^{i+1} in (PI.2) where H_γ^{i+1} may be set-valued. We then evaluate the cost induced by h_γ^{i+1}, namely $V_\gamma^{i+1}(x) = J_\gamma(x, h_\gamma^{i+1})$ for any $x \in \mathbb{R}^{n_x}$, at the evaluation step in (PI.3). By doing so repeatedly, V_γ^i converges to the optimal value function $V_\gamma^\infty = V_\gamma^\star$ under mild conditions, see, e.g., [2].

It is implicitly assumed here that the optimization problem defined in (PI.2) always admits a solution, i.e., $H_\gamma^i(x)$ is non-empty for any $x \in \mathbb{R}^{n_x}$ at any iteration $i \in \mathbb{Z}_{>0}$. We say in this case that Algorithm 1 is *recursively feasible*. We will go back to this point in Section IV-B.

B. Objectives

As we cannot iterate Algorithm 1 infinitely many times, our objective is to give conditions under which PI generates stabilizing policies after a finite number of iterations. We also aim at providing near-optimality guarantees for PI. In

Algorithm 1: Policy Iteration

Input: f in (1), ℓ in (2), $\gamma \in (0,1)$, initial policy $h^0 \in \mathcal{U}$

Output: Policy h_γ^∞, cost V_γ^∞

Initial evaluation step: for all $x \in \mathbb{R}^{n_x}$

$$V_\gamma^0(x) := J_\gamma(x, h^0). \tag{PI.1}$$

for $i \in \mathbb{Z}_{\geq 0}$ **do**

  Policy improvement step: for all $x \in \mathbb{R}^{n_x}$

$$H_\gamma^{i+1}(x) := \underset{u \in \mathcal{U}(x)}{\operatorname{argmin}} \{\ell(x,u) + \gamma V_\gamma^i(f(x,u))\}. \tag{PI.2}$$

  Select $h_\gamma^{i+1} \in H_\gamma^{i+1}$.

  Policy evaluation step: for all $x \in \mathbb{R}^{n_x}$,

$$V_\gamma^{i+1}(x) := J_\gamma(x, h_\gamma^{i+1}). \tag{PI.3}$$

end for

return $h_\gamma^\infty \in H_\gamma^\infty$ and V_γ^∞.

particular, we will see that the bound on $V_\gamma^i - V_\gamma^\star$ we provide significantly differ and improve those encountered in the dynamic programming literature [2,16] for the considered class of systems.

We need to make several assumptions to meet these objectives, which are presented in the next section.

IV. Standing Assumptions

A. Existence of an Optimal Sequence

As mentioned in Section II-B, we assume that for any $x \in \mathbb{R}^{n_x}$, there exists (at least) one infinite-length sequence of admissible inputs minimizing (2).

Standing Assumption 1 (SA1): For any $x \in \mathbb{R}^{n_x}$ and any $\gamma \in (0,1)$, there exists an optimal sequence of admissible inputs $\boldsymbol{u}_\gamma^\star(x)$ such that $V_\gamma^\star(x) = J_\gamma\big(x, \boldsymbol{u}_\gamma^\star(x)\big) < +\infty$ and for any infinite-length sequence of admissible inputs $\boldsymbol{u}$, $V_\gamma^\star(x) \leq J_\gamma(x, \boldsymbol{u})$. □

Condition on system (1) and cost function (2) ensuring SA1 are available in [12] for instance. SA1 ensures the existence of the optimal value function $V_\gamma^\star$ given in (3) as well as the non-emptiness of $H_\gamma^\star(x)$ in (5) for any $x \in \mathbb{R}^{n_x}$ and $\gamma \in (0,1)$. SA1 is very reasonable in the context of this work as we aim to use PI to generate policies, whose costs converge to the optimal one.

B. Recursive Feasibility of PI

We proceed as is often done in the literature, see, e.g., [10,14,15], and assume Algorithm 1 is recursively feasible, in the sense that the set $H_\gamma^i(x)$ is non-empty at any iteration $i \in \mathbb{Z}_{>0}$ and for any $x \in \mathbb{R}^{n_x}$, which is equivalent to say that the optimization problem in (PI.2) admits a solution for any $x \in \mathbb{R}^{n_x}$ at any iteration $i \in \mathbb{Z}_{>0}$.

Standing Assumption 2 (SA2): For any $i \in \mathbb{Z}_{>0}$ and $x \in \mathbb{R}^{n_x}$, the set-valued map $H_\gamma^i(x)$ is non-empty. □

SA2 ensures the recursive feasibility of Algorithm 1 by allowing the selection at each iteration of a new policy. As explained in the introduction, if SA2 does not hold, we can use the modified version of PI presented in [8, Section IV] and the forthcoming results apply *mutatis mutandis*.

C. Detectability

To define stability, we use a continuous and radially unbounded function[1] $\sigma : \mathbb{R}^{n_x} \to \mathbb{R}_{\geq 0}$ that serves as a state measure relating the distance of the state to a given attractor where σ vanishes. As explained in [9,17], σ can be defined as $|\cdot|^p$ when studying the stability of the origin, or as $|\cdot|_{\mathcal{A}}^p$, with $p \in \mathbb{Z}_{>0}$ when studying the stability of non-empty set $\mathcal{A} \subseteq \mathbb{R}^{n_x}$. We make the next detectability assumption on system (1) and stage cost ℓ, which is inspired by [9].

Standing Assumption 3 (SA3): There exist a continuous function $W : \mathbb{R}^{n_x} \to \mathbb{R}_{\geq 0}$, $\alpha_W \in \mathcal{K}_\infty$ and $\overline{\alpha}_W : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ continuous, nondecreasing and zero at zero, such that, for any $x \in \mathbb{R}^{n_x}$ and $u \in \mathcal{U}(x)$,

$$\begin{aligned} W(x) &\leq \overline{\alpha}_W\big(\sigma(x)\big) \\ W\big(f(x,u)\big) - W(x) &\leq -\alpha_W\big(\sigma(x)\big) + \ell(x,u). \end{aligned} \tag{7}$$

□

SA3 is a detectability property of system (1) with respect to σ when considering ℓ as an output. This is very natural as this captures the fact, that by minimizing ℓ along the solutions to (1), desirable stability properties should follow. SA3 is consistent with the literature on LQ [1]; the link between (7) and detectability of linear time-invariant systems being established in [17, Lemma 4]. When $\ell(x,u) = q(x) + r(u)$ for any $x \in \mathbb{R}^{n_x}$ and any $u \in \mathcal{U}(x)$, with $q : \mathbb{R}^{n_x} \to \mathbb{R}_{\geq 0}$ and $r : \mathbb{R}^{n_u} \to \mathbb{R}_{\geq 0}$ such that there exists $\underline{\alpha} \in \mathcal{K}_\infty$ verifying $q(x) \geq \underline{\alpha}(\sigma(x))$ for any $x \in \mathbb{R}^{n_x}$, then SA3 is verified with $W = 0$, $\alpha_W = \underline{\alpha}$ and $\overline{\alpha}_W = 0$.

Remark 1: A more general detectability property is considered in [9,17] where the second line of (7) is given by $W(f(x,u)) - W(x) \leq -\alpha_W(\sigma(x)) + \chi(\ell(x,u))$ for any $x \in \mathbb{R}^{n_x}$, $u \in \mathbb{R}^{n_u}$ and with $\chi \in \mathcal{K}_\infty$. We plan to investigate this generalization in future work. □

D. Initial Policy

We also make the next assumption on the cost given by the initial policy.

Standing Assumption 4 (SA4): Let $h^0 \in \mathcal{U}$ be known and such that there exist $\gamma_0 \in (0,1]$ and $\overline{\alpha}_V : \mathbb{R}_{\geq 0} \times (0, \gamma_0) \to \mathbb{R}_{\geq 0}$ of class $\mathcal{K}_\infty$ in its first argument verifying for any $x \in \mathbb{R}^{n_x}$ and any $\gamma \in (0, \gamma_0)$,

$$V_\gamma^0(x) = J_\gamma(x, h^0) \leq \overline{\alpha}_V\big(\sigma(x), \gamma\big). \tag{8}$$

□

SA4 requires the knowledge of an initial policy h^0 such that there exists a range of values for γ, namely $(0, \gamma_0)$, such that the initial cost $V_\gamma^0(x)$ is finite for every $x \in \mathbb{R}^{n_x}$ and is upper-bounded by function $\overline{\alpha}_V$, which is $\mathcal{K}_\infty$ in its first argument and depends on γ. It is important to note that SA4 may hold even if h^0 is not stabilizing as illustrated in

[1] In the sense that for any $\Delta > 0$, $\{x \in \mathbb{R}^{n_x} : \sigma(x) \leq \Delta\}$ is compact.

Section VI and exemplified below. The next lemma gives a sufficient condition to ensure SA4.

Lemma 1: Consider system (1) and suppose there exist $M, a > 0$, $\chi \in \mathcal{K}_\infty$ and an admissible policy $h \in \mathcal{U}$ such that for any $x \in \mathbb{R}^{n_x}$ and any $k \in \mathbb{Z}_{\geq 0}$, $\ell\big(\phi(k,x,h), h(\phi(k,x,h))\big) \leq M a^k \chi(\sigma(x))$. Then SA4 is verified with $h^0 = h$, $\gamma_0 = \min\big\{1, \frac{1}{a}\big\}$ and $\overline{\alpha}_V(s,\gamma) = \frac{M}{1-a\gamma}\chi(s)$ for any $s \in \mathbb{R}_{\geq 0}$ and any $\gamma \in (0, \gamma_0)$. □

Proof. Let $h^0 = h$ and $\gamma \in (0, \gamma_0)$ with h and γ_0 as in Lemma 1. For any $x \in \mathbb{R}^{n_x}$,

$$V_\gamma^0(x) = \sum_{k=0}^{\infty} \gamma^k \ell\big(\phi(k,x,h^0), h(\phi(k,x,h^0))\big) \leq M\chi(\sigma(x)) \sum_{k=0}^{\infty} (a\gamma)^k. \tag{9}$$

As $\gamma < \gamma_0 \leq \frac{1}{a}$, we have $V_\gamma^0(x) \leq \frac{M}{1-a\gamma}\chi(\sigma(x))$. Thus SA4 holds with $\overline{\alpha}_V(\cdot,\gamma) := \frac{M}{1-a\gamma}\chi$ and this concludes the proof. ■

Lemma 1 shows that, if the stage cost along the solution to (1) with policy h^0 is upper-bounded at any time-step $k \in \mathbb{Z}_{\geq 0}$ by a term $Ma^k\sigma(x)$, we can determine explicitly γ_0 and $\overline{\alpha}_V(\cdot,\gamma)$ verifying SA4. Note that a in Lemma 1 may be strictly bigger than 1, which implies that h^0 may not be stabilizing.

Remark 2: It is sometimes difficult to determine a stabilizing initial policy for general nonlinear systems. Hence the main idea is to remove this constraint by exploiting the discount factor. □

E. Stability with Optimal Sequence

As we want to eventually obtain stabilizing policies using PI, we will assume that optimal policies are stabilizing. The next assumption together with SA1 and SA3 indeed guarantee that the closed-loop system (1) with optimal controller, i.e., system (6), verifies a $\mathcal{KL}$-stability property with respect to σ as established in Proposition 1 below.

Standing Assumption 5 (SA5): The following holds.

(i) There exists $\overline{\alpha}_{V^\star} \in \mathcal{K}_\infty$ such that for any $\gamma \in (0, \gamma_0)$ with γ_0 in SA4, for any $x \in \mathbb{R}^{n_x}$, $V_\gamma^\star(x) \leq \overline{\alpha}_{V^\star}\big(\sigma(x)\big)$ where $V_\gamma^\star$ is the optimal value function in (3).

(ii) There exists $\gamma^\star \in (0, \gamma_0)$ such that

$$(1-\gamma^\star)\overline{\alpha}_{V^\star}(s) \leq \alpha_W(s), \qquad \forall s \in \mathbb{R}_{>0}, \tag{10}$$

with α_W in SA3. □

Item (i) of SA5 holds for instance when $\overline{\alpha}_V$ in (8) is independent of γ as $V_\gamma^\star(x) \leq V_{\gamma_0}^\star(x)$ for any $x \in \mathbb{R}^{n_x}$ and any $\gamma \in (0, \gamma_0)$. It is important to note that we do not need to know either $V_\gamma^\star$ or $V_{\gamma_0}^\star$ to check whether item (i) of SA5 holds. Indeed, this condition holds whenever the cost for a known, not necessarily optimal, policy is upper-bounded by $\overline{\alpha}_{V^\star}(\sigma(x))$ for any $x \in \mathbb{R}^{n_x}$ for some $\overline{\alpha}_{V^\star} \in \mathcal{K}_\infty$. A condition ensuring item (i) is given in [17, Lemma 1]. Item (ii) of SA5 is a technical condition useful to deduce global asymptotic stability properties for system (6) as formalized next.

Proposition 1: For any $\gamma \in (\gamma^\star, \gamma_0)$, system (6) is *$\mathcal{KL}$-stable with respect to σ*, i.e., there exists $\beta_\gamma^\star \in \mathcal{KL}$ such that for any $x \in \mathbb{R}^{n_x}$, any solution $\phi_\gamma^\star(\cdot, x)$ to system (6) satisfies

$$\sigma\big(\phi_\gamma^\star(k,x)\big) \leq \beta_\gamma^\star\big(\sigma(x), k\big) \quad \forall k \in \mathbb{Z}_{\geq 0}. \tag{11}$$

In particular, $\beta_\gamma^\star : (s,k) \mapsto \underline{\alpha}_{Y^\star}^{-1}(\widetilde{\beta}_\gamma^\star(\overline{\alpha}_{Y^\star}(s), k)) \in \mathcal{KL}$ with $\underline{\alpha}_{Y^\star}$, $\widetilde{\beta}_\gamma^\star$ and $\overline{\alpha}_{Y^\star}$ in Table 1. □

Remark 3: Proposition 1 ensures a global asymptotic stability property for system (6). We will investigate in future work the case where a semiglobal practical stability holds for (6) instead as in [17], which will allow us to relax item (ii) of SA5. □

Now that all the standing assumptions have been stated, we are ready to present the main results.

V. Main results

In this section, we consider system (1) whose inputs are generated by PI at iteration $i \in \mathbb{Z}_{\geq 0}$, that is,

$$x(k+1) \in f\big(x(k), H_\gamma^i(x(k))\big) =: F_\gamma^i\big(x(k)\big), \quad \forall i \in \mathbb{Z}_{\geq 0}. \tag{12}$$

For convenience, solutions to system (12) are denoted in the sequel as $\phi_\gamma^i(k,x)$ when initialized at $x \in \mathbb{R}^{n_x}$ for any $k \in \mathbb{Z}_{\geq 0}$.

A. Near-optimality

We first recover the classical result presented in, e.g., [3], on the improvement property of the policies generated by PI, whose proof is omitted as it follows similar lines as [10, Lemma 2].

Lemma 2: For any $x \in \mathbb{R}^{n_x}$, $i \in \mathbb{Z}_{\geq 0}$ and $\gamma \in (0, \gamma_0)$ with γ_0 in SA4, $V_\gamma^{i+1}(x) \leq V_\gamma^i(x)$. □

In the next theorem, we establish a new near-optimality bound for PI with discounted costs.

Theorem 1: For any $i \in \mathbb{Z}_{\geq 0}$, $x \in \mathbb{R}^{n_x}$, $\gamma \in \big(\gamma^\star, \gamma_0\big)$ and any solution $\phi_\gamma^\star(\cdot, x)$ to system (6),

$$\begin{aligned}\big(V_\gamma^i - V_\gamma^\star\big)(x) &\leq \gamma^i\big(V_\gamma^0 - V_\gamma^\star\big)(\phi_\gamma^\star(i,x)) \\ &\leq \gamma^i \overline{\alpha}_V(\beta_\gamma^\star(\sigma(x), i), \gamma),\end{aligned} \tag{13}$$

with $\beta_\gamma^\star \in \mathcal{KL}$ from Proposition 1 and $\overline{\alpha}_V$ from SA4. □

Theorem 1 gives us an explicit upper-bound on the term $V_\gamma^i(x) - V_\gamma^\star(x)$ for any $x \in \mathbb{R}^{n_x}$ and any iteration $i \in \mathbb{Z}_{\geq 0}$. Hence, this bound provides an estimation on how close cost V_γ^i generated at iteration $i \in \mathbb{Z}_{\geq 0}$ by PI is from the "target" $V_\gamma^\star$. As $\beta_\gamma^\star \in \mathcal{KL}$ for any $\gamma \in (\gamma^\star, \gamma_0)$, the upper bound in (13) converges to 0 as the number of iteration i goes to infinity. Typical near-optimal bounds for PI in the literature are of the form $\frac{M_1\gamma^i}{1-\gamma} + M_2$ with $M_1, M_2 \in \mathbb{R}_{>0}$, see, e.g., [2,16]. The latter bound explodes as γ goes to 1 and does not vanish to 0 as $\sigma(x)$ goes to 0, contrary to the bound in Theorem 1. This can be explained by the fact that the bounds

of the literature do not exploit the stability properties of the optimal policies established in Proposition 1.

We now focus on the stability guarantees that can be deduced thanks to this near-optimality property.

B. Stability

We first establish a Lyapunov property for system (12) for any iteration $i \in \mathbb{Z}_{\geq 0}$.

Theorem 2: There exist $\underline{\alpha}_Y \in \mathcal{K}_\infty$, $\overline{\alpha}_Y$, $\alpha_Y : \mathbb{R}_{\geq 0} \times (\gamma^\star, \gamma_0) \to \mathbb{R}_{\geq 0}$ of class $\mathcal{K}_\infty$ in their first argument such that for any $i \in \mathbb{Z}_{\geq 0}$ there exist $Y_\gamma^i : \mathbb{R}^{n_x} \to \mathbb{R}_{\geq 0}$ and $\Upsilon^i : \mathbb{R}_{\geq 0} \times (\gamma^\star, \gamma_0) \to \mathbb{R}_{\geq 0}$ of class $\mathcal{K}_\infty$ in its first argument such that the following holds for any $\gamma \in (\gamma^\star, \gamma_0)$.

(i) For any $x \in \mathbb{R}^{n_x}$, $\underline{\alpha}_Y\big(\sigma(x)\big) \leq Y_\gamma^i(x) \leq \overline{\alpha}_Y\big(\sigma(x), \gamma\big)$.

(ii) For any $x \in \mathbb{R}^{n_x}$ and $v \in F_\gamma^i(x)$,

$$Y_\gamma^i(v) - Y_\gamma^i(x) \leq \tfrac{1}{\gamma}\Big(- \alpha_Y\big(\sigma(x), \gamma\big) + \Upsilon^i\big(\sigma(x), \gamma\big)\Big),$$

where $\underline{\alpha}_Y$, $\overline{\alpha}_Y$, α_Y, Υ^i and Y_γ^i are defined in Table I. □

Item (i) of Theorem 2 means that Y_γ^i is positive definite and radially unbounded with respect to σ for any $\gamma \in (\gamma^\star, \gamma_0)$. Item (ii) of Theorem 2 is a dissipative inequality of system (12) for which the supply rate consists of a negative term, namely $-\frac{1}{\gamma}\alpha_Y(\cdot, \gamma)$, and a non-negative term $\frac{1}{\gamma}\Upsilon^i(\cdot, \gamma)$ which can be made as small as desired by increasing i. The latter property is key for establishing stability properties for system (12) with i sufficiently large.

$\underline{\alpha}_{Y^\star}$, $\underline{\alpha}_Y$	α_W
$\alpha_{Y^\star}(\cdot, \gamma)$, $\alpha_Y(\cdot, \gamma)$	$\frac{\gamma - \gamma^\star}{1 - \gamma^\star}\alpha_W$
$\overline{\alpha}_{Y^\star}$	$\overline{\alpha}_{V^\star} + \frac{1}{\gamma^\star}\overline{\alpha}_W$
$\widetilde{\alpha}_{Y^\star}(\cdot, \gamma)$	$\alpha_Y(\cdot, \gamma) \circ \overline{\alpha}_{Y^\star}$
$\widetilde{\beta}_\gamma^\star(s, k)$	$\max_{\widehat{s} \in [0, s]} \big(\mathbb{I} - \frac{1}{\gamma}\widetilde{\alpha}_{Y^\star}(\cdot, \gamma)\big)^{(k)}(\widehat{s})$
Y_γ^i	$V_\gamma^i + \frac{1}{\gamma}W$
$\overline{\alpha}_Y(\cdot, \gamma)$	$\overline{\alpha}_V(\cdot, \gamma) + \frac{1}{\gamma}\overline{\alpha}_W$
Υ^i	$(s_1, s_2) \mapsto (1 - s_2)s_2^i \overline{\alpha}_V\big(\beta_\gamma^\star(s_1, i), s_2\big)$

TABLE I: Expressions of functions used in Theorems 1 and 2

Once all the functions in SA3 and SA4 are identified, the functions appearing in Table I may be derived explicitly, see Section VI for examples.

Based on Theorem 2, we establish the next stability property for system (6).

Theorem 3: For any $\gamma \in (\gamma^\star, \gamma_0)$, there exists $\beta_\gamma \in \mathcal{KL}$ such that for any $\delta, \Delta > 0$, $i \geq i_\gamma^\star$ with $i_\gamma^\star \in \mathbb{Z}_{\geq 0}$ verifying

$$i_\gamma^\star \geq \frac{\ln\Big(\frac{\alpha_Y(\overline{\alpha}_Y^{-1}(\underline{\alpha}_Y(\delta), \gamma), \gamma)}{2(1-\gamma)\overline{\alpha}_V(\beta_\gamma^\star(\underline{\alpha}_Y^{-1}(\overline{\alpha}_Y(\Delta, \gamma)), 0), \gamma)}\Big)}{\ln(\gamma)}, \tag{14}$$

any $x \in \{z \in \mathbb{R}^{n_x} : \sigma(z) \leq \Delta\}$, any solution $\phi_\gamma^i(\cdot, x)$ to system (12) satisfies

$$\sigma\big(\phi_\gamma^i(k, x)\big) \leq \max\{\beta_\gamma\big(\sigma(x), k\big), \delta\} \quad \forall k \in \mathbb{Z}_{\geq 0}. \tag{15}$$

□

Theorem 3 ensures a semiglobal and pratical stability property of (12) for any $\gamma \in (\gamma^\star, \gamma_0)$ where the tuning parameter is the number of iterations i. In particular, for any set of initial conditions of the form $\{x \in \mathbb{R}^{n_x} : \sigma(x) \leq \Delta\}$ where $\Delta > 0$ can be arbitrarily large, for any (arbitrarily small) $\delta > 0$, we can always compute $i_\gamma^\star \in \mathbb{Z}_{\geq 0}$ verifying (14) such that for any iteration $i \geq i_\gamma^\star$ (15) holds. By strengthening the conditions of Theorem 3, it is possible to derive stronger stability guarantees.

Corollary 1: Suppose there exist $\overline{a}_W \geq 0$, $a_W, \overline{a}_{V^\star} > 0$ and $\overline{a}_V : (\gamma^\star, \gamma_0) \mapsto \mathbb{R}_{>0}$ such that $\overline{\alpha}_V(s, \cdot) = \overline{a}_V(\cdot)s$, $\overline{\alpha}_{V^\star}(s) = \overline{a}_{V^\star}s$, $\underline{\alpha}_W(s) = a_W s$ and $\overline{\alpha}_W(s) \leq \overline{a}_W s$ for any $s \geq 0$ with $\gamma^\star = \frac{\overline{a}_{V^\star} - a_W}{\overline{a}_{V^\star}}$. Then for any $\gamma \in \big(\gamma^\star, \gamma_0\big)$ there exists $i_\gamma^\star \in \mathbb{Z}_{\geq 0}$ verifying

$$i_\gamma^\star \geq \frac{\ln\Big(\frac{\gamma^\star(\gamma - \gamma^\star)a_W^2}{2\gamma(1-\gamma)^2\overline{a}_V(\gamma)(\gamma^\star\overline{a}_{V^\star} + \overline{a}_W)}\Big)}{\ln\Big(\gamma - \frac{\gamma^\star(\gamma - \gamma^\star)a_W}{(1-\gamma)(\gamma^\star\overline{a}_{V^\star} + \overline{a}_W)}\Big)}, \tag{16}$$

such that for any $i \geq i_\gamma^\star$, $x \in \mathbb{R}^{n_x}$ any solution $\phi_\gamma^i(\cdot, x)$ to system (12) satisfies

$$\sigma\big(\phi_\gamma^i(k, x)\big) \leq c_1(\gamma)\sigma(x)e^{-c_2(\gamma)k} \quad \forall k \in \mathbb{Z}_{\geq 0} \tag{17}$$

with $c_1(\gamma) = \frac{\gamma\overline{a}_V(\gamma) + \overline{a}_W}{\gamma a_W}$ and $c_2(\gamma) = -\ln\big(1 - \frac{a_W(\gamma - \gamma^\star)}{2\gamma(1-\gamma)(\overline{a}_V(\gamma) + \frac{1}{\gamma}\overline{a}_W)}\big) > 0$. □

Corollary 1 ensures that after a sufficient number of iterations $i_\gamma^\star$, that we can explicitly estimate using (16), a global exponential stability property of (12) is also verified for any $\gamma \in (\gamma^\star, \gamma_0)$.

Remark 4: Corollary 1 also ensures a global exponential stability property of the system (6) with a lower bound for $\gamma^\star$ less conservative than those presented in [17, Corollary 2] and [6, Lemma 2]. Indeed $\gamma^\star = \frac{\overline{a}_{V^\star} - a_W}{\overline{a}_{V^\star}} = 1 - \frac{a_W}{\overline{a}_{V^\star}} \leq 1 - \frac{a_W}{\overline{a}_{V^\star} + \overline{a}_W} =: \gamma_{[6]}^\star$ and as $1 > \frac{(\overline{a}_{V^\star} - a_W)}{\overline{a}_{V^\star}}\frac{(\overline{a}_{V^\star} + a_W)}{\overline{a}_{V^\star}} = \frac{\overline{a}_{V^\star}^2 - a_W^2}{\overline{a}_{V^\star}^2}$, we have $\gamma^\star = \frac{\overline{a}_{V^\star} - a_W}{\overline{a}_{V^\star}} \leq \frac{\overline{a}_{V^\star}}{\overline{a}_{V^\star} + a_W} =: \gamma_{[17]}^\star$. □

VI. Examples

We consider two examples, namely the linear quadratic problem and a nonholonomic integrator, for which we show that the standing assumptions hold thereby implying that the results of Section V apply.

A. Linear Quadratic Problem

We consider the deterministic linear time-invariant system

$$x(k+1) = Ax(k) + Bu(k), \tag{18}$$

where $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$ and (A, B) stabilizable. Let $\sigma(x) = |x|^2$ and $\ell(x, u) = x^\top Q x + u^\top R u$ for any $x \in \mathbb{R}^{n_x}$ and $u \in \mathbb{R}^{n_u}$, where $Q = C^\top C \in \mathbb{R}^{n_x \times n_x}$ with (A, C) detectable, and $R \in \mathbb{R}^{n_u \times n_u}$ is a symmetric, positive definite matrix. We set $\mathcal{U}(x) = \mathbb{R}^{n_u}$ for any $x \in \mathbb{R}^{n_x}$.

First, as (A, B) is stabilizable and (A, C) is detectable SA1 holds by [2]. In addition, SA3 is verified with $W(x) = x^\top S_2 x$ for any $x \in \mathbb{R}^{n_x}$, $\alpha_W(s) = \lambda_{\min}(S_1)s$ and $\overline{\alpha}_W(s) = \lambda_{\max}(S_2)s$ for any $s \in \mathbb{R}_{\geq 0}$, where S_1, S_2 are symmetric positive definite matrices satisfying

$$\begin{pmatrix} A^\top S_2 A - S_2 + S_1 - Q & A^\top S_2 B \\ B^\top S_2 A & B^\top S_2 B - R \end{pmatrix} \leq 0. \tag{19}$$

Note that there always exist such matrices S_1 and S_2 by [17, Lemma 4]. Furthermore, the conditions of Lemma 1 are satisfied by taking $h(x) = K_0 x$ for any $x \in \mathbb{R}^{n_x}$, with any $K_0 \in \mathbb{R}^{n_u \times n_x}$, $\chi = \mathbb{I}$, $M = \left\|Q + K_0^\top R K_0\right\|$ and $a = \|A + BK_0\|^2$. Hence SA4 is verified with $\gamma_0 = \min\left\{1, \frac{1}{\|A+BK_0\|^2}\right\}$ and $\overline{\alpha}_V(s,\gamma) := \frac{\left\|Q+K_0^\top R K_0\right\|}{1-\gamma\|A+BK_0\|^2} s$ for any $s \in \mathbb{R}_{\geq 0}$ and $\gamma \in (0, \gamma_0)$. We note that SA4 is verified for any $K_0 \in \mathbb{R}^{n_u \times n_x}$, thus even when $A + BK_0$ is not Schur, i.e., even when the initial policy is not stabilizing. Moreover, using the same time-varying change of coordinates as in [13], we know by [14] that SA2 holds, as $\sqrt{\gamma}(A + BK_0)$ is Schur for any $\gamma < \gamma_0$ (which does not mean $A + BK_0$ is Schur obviously). Given (A, B) stabilizable, the optimal value function for any $\gamma \in [0, 1]$ and $x \in \mathbb{R}^{n_x}$ is $V_\gamma^\star(x) := x^\top P_\gamma^\star x$, where $P_\gamma^\star$ is a symmetric matrix. Hence SA5 holds with $\overline{\alpha}_{V^\star}(s) = \lambda_{\max}(P_1^\star)s$ for any $s \geq 0$ and $\gamma^\star = 1 - \frac{\lambda_{\min}(S_1)}{\lambda_{\max}(P_1^\star)}$. Provided we take K_0 such that $\gamma_0 > \gamma^\star$, all the conditions of Corollary 1 are verified and we conclude that the system (12) verifies a global exponential stability property for any $\gamma \in (\gamma^\star, \gamma_0)$ after a sufficient number of iterations whose estimate is given by (16).

B. Nonholonomic Integrator

Consider the nonholonomic integrator as in [9, Example 2], that is,

$$x_1^+ = x_1 + u_1 \quad x_2^+ = x_2 + u_2 \quad x_3^+ = x_3 + x_1 u_2 - x_2 u_1, \quad (20)$$

where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, $u = (u_1, u_2) \in \mathcal{U}(x) = \mathbb{R}^2$. Let $\sigma(x) = x_1^2 + x_2^2 + 10\,|x_3|$ and $\ell(x, u) = \sigma(x) + |u|^2$ for any $x \in \mathbb{R}^3$ and $u \in \mathbb{R}^2$.

Thanks to [17], SA1, SA3 and item (i) of SA5 are verified with $W = 0$, $\underline{\alpha}_W = 0$, $\alpha_W = \mathbb{I}$ and $\alpha_{V^\star} = \frac{22}{5}\mathbb{I}$. We assume that SA2 holds. As the conditions of Lemma 1 are satisfied by taking $h(x) = (\frac{1}{15}x_1, -x_2)$ for any $x \in \mathbb{R}^{n_x}$ with $\chi = \mathbb{I}$, $a = \frac{256}{225}$, $M = \frac{22}{3}$. As a consequence, SA4 is satisfied with $\gamma_0 = \frac{225}{256}$ and $\alpha_V(\cdot, \gamma) = \frac{M}{1-a\gamma}\mathbb{I}$ for any $\gamma \in (0, \gamma_0)$. Moreover, using the condition given in Corollary 1, item (ii) of SA5 is ensured with $\gamma^\star = \frac{17}{22}$. As $\frac{17}{22} \approx 0.77 < 0.88 \approx \frac{225}{256}$, we have $\gamma^\star < \gamma_0$ such that all the conditions of Corollary 1 are verified. We conclude that the system (12) verifies a global exponential stability property for any $\gamma \in (\gamma^\star, \gamma_0)$ after a sufficient number of iterations whose estimate is given by $i_\gamma^\star = \left\lceil \frac{\ln\left(\frac{(330\gamma-255)(1-\frac{256}{225}\gamma)}{21296\gamma(1-\gamma)^2}\right)}{\ln\left(\gamma - \frac{110\gamma-85}{484(1-\gamma)}\right)} \right\rceil$. Thus, when $\gamma = 0.86$, $i_\gamma^\star = 20$, which means that the systems controlled by the policies generated by PI exhibit an exponential stability property after 20 iterations.

VII. Conclusion

We have analyzed the stability of general nonlinear discrete-time systems controlled by sequences of inputs generated by PI for an infinite-horizon discounted cost. Inspired by [8], novel near-optimal bounds have been established for the discounted problem. These new bounds do not blow up when the discount factor tends to 1 contrary to, e.g., [2,16]. The novel near-optimality bounds were then exploited to also provide general conditions under which PI generates stabilizing policies after sufficiently many iterations.

In future work, we plan to relax some of the made assumptions in particular SA3 and item (ii) of SA5, and to investigate the case where γ is increased with the number of iterations as done in [13] for the LQ problem.

References

[1] B.D.O. Anderson and J.B. Moore. *Optimal Control: Linear Quadratic Methods*. Dover edition, Mineola, U.S.A., 2007.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, U.S.A., 4th edition, 2012.

[3] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, U.S.A., 1996.

[4] J. de Brusse, M. Granzotto, R. Postoyan, and D. Nešić. Policy iteration for discrete-time systems with discounted costs: stability and near-optimality guarantees. *arXiv preprint arXiv:2403.19007*, 2024.

[5] R. Goebel, R.G. Sanfelice, and A.R. Teel. *Hybrid Dynamical Systems*. Princeton University Press, Princeton, U.S.A., 2012.

[6] M. Granzotto, R. Postoyan, L. Buşoniu, D. Nešić, and J. Daafouz. Finite-horizon discounted optimal control: stability and performance. *IEEE Transactions on Automatic Control*, 66(2):550–565, 2020.

[7] M. Granzotto, R. Postoyan, D. Nešić, L. Buşoniu, and J. Daafouz. When to stop value iteration: stability and near-optimality versus computation. In *3rd Conference on Learning for Dynamics and Control, PMLR 144:412-424*, 2021.

[8] M. Granzotto, O. Lindamulage De Silva, R. Postoyan, D. Nešić, and Z.-P. Jiang. Robust stability and near-optimality for policy iteration: for want of recursive feasibility, all is not lost. *IEEE Transactions on Automatic Control, available on-line*, 2024.

[9] G. Grimm, M.J. Messina, S.E. Tuna, and A.R. Teel. Model predictive control: for want of a local control Lyapunov function, all is not lost. *IEEE Transactions on Automatic Control*, 50(5):546–558, 2005.

[10] A. Heydari. Analyzing policy iteration in optimal control. In *American Control Conference, Boston, U.S.A.*, pages 5728–5733, 2016.

[11] Y. Jiang and Z.-P. Jiang. *Robust Adaptive Dynamic Programming*. Wiley-IEEE Press, 2017.

[12] S.S. Keerthi and E.G. Gilbert. An existence theorem for discrete-time infinite-horizon optimal control problems. *IEEE Transactions on Automatic Control*, 30(9):907–909, 1985.

[13] A. Lamperski. Computing stabilizing linear controllers via policy iteration. In *IEEE Conference on Decision and Control, Jeju Island: South Korea*, pages 1902–1907, 2020.

[14] F. L. Lewis and D. Vrabie. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Transactions on Automatic Control*, 9(3):32–50, 2009.

[15] D. Liu and Q. Wei. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 25(3):621–634, 2013.

[16] R. Munos. Error bounds for approximate policy iteration. *ICML*, 3:560–567, 2003.

[17] 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*, 62(6):2736–2749, 2017.

[18] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, U.S.A., 1998.

[19] H. Zhang, D. Liu, Y. Luo, and D. Wang. *Adaptive Dynamic Programming for Control: Algorithms and Stability*. Springer, 2012.