Proceedings of Machine Learning Research vol 144:1–13, 2021

When to stop value iteration: stability and near-optimality versus computation

Mathieu Granzotto[†] MATHIEU.GRANZOTTO@UNIV-LORRAINE.FR
[†]*Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France.*

Romain Postoyan[†] ROMAIN.POSTOYAN@UNIV-LORRAINE.FR

Dragan Nešić DNESIC@UNIMELB.EDU.AU
Electrical and Electronic Engineering Department, University of Melbourne, Parkville, VIC 3010, Australia.

Lucian Buşoniu LUCIAN.BUSONIU@AUT.UTCLUJ.RO
Department of Automation, Technical University of Cluj-Napoca, Memorandumului 28, 400114, Romania.

Jamal Daafouz[†] JAMAL.DAAFOUZ@UNIV-LORRAINE.FR

Abstract

Value iteration (VI) is a ubiquitous algorithm for optimal control, planning, and reinforcement learning schemes. Under the right assumptions, VI is a vital tool to generate inputs with desirable properties for the controlled system, like optimality and Lyapunov stability. As VI usually requires an infinite number of iterations to solve general nonlinear optimal control problems, a key question is when to terminate the algorithm to produce a "good" solution, with a measurable impact on optimality and stability guarantees. By carefully analysing VI under general stabilizability and detectability properties, we provide explicit and novel relationships of the stopping criterion's impact on near-optimality, stability and performance, thus allowing to tune these desirable properties against the induced computational cost. The considered class of stopping criteria encompasses those encountered in the control, dynamic programming and reinforcement learning literature and it allows considering new ones, which may be useful to further reduce the computational cost while endowing and satisfying stability and near-optimality properties. We therefore lay a foundation to endow machine learning schemes based on VI with stability and performance guarantees, while reducing computational complexity.

1. Introduction

Value iteration (VI) is an established method for optimal control, which plays a key role in reinforcement learning (Sutton and Barto, 2017; Lewis and Vrabie, 2009; Buşoniu et al., 2018; Pang et al., 2019). This algorithm consists in iteratively constructing approximations of the optimal value function, based on which near-optimal control inputs are derived for a given dynamical nonlinear systems and a given stage cost. The convergence of said approximations to the optimal value function is established in, e.g., (Bertsekas, 2012, 2017) under mild conditions. To benefit from this convergence property, VI often needs to be iterated infinitely many times. However, in practice, we cannot do so and must stop iterating the algorithm before to manage the computational burden, which may be critical in online applications. Heuristics are often used in the literature to stop iterating by comparing the mismatch between the value functions obtained at the current step and at the previous one, see, e.g., (Bertsekas, 2012; Sutton and Barto, 2017; Pang et al., 2019; Kiumarsi et al., 2017; Liu et al., 2015). An important question is then how far the obtained approximate value function is to the optimal one. To the best of our knowledge, this is only analysed in general when the cost is

discounted and the stage cost takes values in a bounded set (Bertsekas, 2012). An alternative consists in asking for a sufficiently large number of iterations, as the near-optimality gap vanishes as the number of iterations increases, e.g. (Bertsekas, 2012; Heydari, 2018, 2014, 2016; Liu et al., 2015; Granzotto et al., 2020a), but the issue is then the computational cost. Indeed, any estimate of the number of iterations is in general subject to conservatism, and, as a result, we may iterate many more times than what is truly required to ensure "good" near-optimality properties. There is therefore a need for stopping criteria for VI whose impact on near-optimality is analytically established, and which are not too computationally demanding.

Our main goal is to use VI to simultaneously ensure near-optimal control and stability properties for physical systems. Stability is critical in many applications, as: (i) it provides analytical guarantees on the behavior of the controlled system solutions as time evolves; (ii) endows robustness properties and is thus associated to safety considerations, see, e.g., (Berkenkamp et al., 2017). We therefore consider systems and costs where general stability properties are bestowed by VI based schemes, which follows from assumed general stabilizability and detectability properties of the plant model and the stage cost as in (Grimm et al., 2005; Postoyan et al., 2017; Granzotto et al., 2020a).

In this context, we consider state-dependent stopping criteria for VI and we analyse their impact on the near-optimality and stability properties of the obtained policies for general deterministic nonlinear plant models and stage costs, where no discount factor is employed. Instead of relying on a uniform contraction property as in, e.g., (Bertsekas, 2012; Liu et al., 2015), our analysis is centered on and exploits stabilizability and detectability properties of the plant and stage costs, which are expressed in terms of Lyapunov inequalities. Our work covers the state-independent stopping criteria considered in the control, dynamical programming and reinforcement learning literature (Sutton and Barto, 2017; Lewis and Vrabie, 2009; Buşoniu et al., 2018), but provides analytical guarantees for undiscounted stage costs taking values in unbounded sets. By carefully analysing the stopping criterion's impact on near-optimality, stability and closed-loop cost guarantees, we provide means to tune these properties against the induced computational cost, thus clarifying the tradeoff between "good enough" convergence of VI and "good properties" of generated inputs. Considering that VI is, via Q-learning, the basis of many state-of-the-art reinforcement learning methods, we believe the results of this paper contribute to the (near)-optimality analysis for reinforcement learning, as we lay a foundation to endow such schemes with stability and performance guarantees, while reducing computational complexity.

The paper and its contributions are organized as follows. In Section 2, we formally state the problem and the main assumptions. We introduce the design of stopping criteria for VI in Section 3, and show that the VI stopping criterion is indeed verified with a finite number of iterations. Our main results are found in Section 4. There, we provide near-optimal guarantees, i.e. a bound on the mismatch between the approximated value function and the true optimal value function. The bound can be easily and directly tuned by the designed stopping criterion. Additionally, stability and performance guarantees of the closed-loop system with inputs generated by VI are provided, given that the stopping criterion is appropriately chosen. In Section 5, we provide an example to illustrate our results. Concluding remarks are drawn in Section 6. The proofs are omitted and available in the associated technical report (Granzotto et al., 2020c).

Prior literature. The classical stopping criterion is analysed in (Bertsekas, 2012), albeit restricted to when the cost is discounted and the stage cost takes values in a bounded set. Concerning stability, works like (Granzotto et al., 2020a; Heydari, 2017; Wei et al., 2015) provide conditions to ensure that the feedback law obtained ensures a stability property for a dynamical system. In particular,

it is required in (Granzotto et al., 2020a) that the number of iteration d be sufficiently large, and lower bounds on d are provided, but these are subject to some conservatism. As explained above, by adapting the number of iterations with data available during computations, the algorithm avoids the conservatism often incurred by offline estimations for stability and near-optimality guarantees. This is indeed the case in an example (see Section 5), where we observe 91% fewer iterations for comparable guarantees. Similar ideas related to the stopping criterion were exploited in (Granzotto et al., 2020b), for a different purpose, namely for the redesign of optimistic planning (Hren and Munos, 2008) to address the near-optimal control of switched systems. We are also aware of work of (Pavlov et al., 2019), which adapts the stopping criterion with stability considerations for interior point solvers for reduced computational complexity for nonlinear model predictive control applications.

Notation. Let $\mathbb{R} := (-\infty, \infty)$, $\mathbb{R}_{\geq 0} := [0, \infty)$, $\mathbb{Z}_{\geq 0} := \{0, 1, 2, \ldots\}$ and $\mathbb{Z}_{>0} := \{1, 2, \ldots\}$. We use (x, y) to denote $[x^\top, y^\top]^\top$, where $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ and $n, m \in \mathbb{Z}_{>0}$. A function $\chi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of class $\mathcal{K}$ if it is continuous, zero at zero and strictly increasing, and it is of class $\mathcal{K}_\infty$ if it is of class $\mathcal{K}$ and unbounded. A continuous function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of class $\mathcal{KL}$ when $\beta(\cdot, t)$ is of class $\mathcal{K}$ for any $t \geq 0$ and $\beta(s, \cdot)$ is decreasing to 0 for any $s \geq 0$. The notation $\mathbb{I}$ stands for the identity map from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$. For any sequence $\boldsymbol{u} = [u_0, u_1, \ldots]$ of length $d \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$ where $u_i \in \mathbb{R}^m$, $i \in \{0, \ldots, d\}$, and any $k \in \{0, \ldots, d\}$, we use $\boldsymbol{u}|_k$ to denote the first k elements of $\boldsymbol{u}$, i.e. $\boldsymbol{u}|_k = [u_0, \ldots, u_{k-1}]$ and $\boldsymbol{u}|_0 = \varnothing$ by convention. Let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, we use $g^{(k)}$ for the composition of function g with itself k times, where $k \in \mathbb{Z}_{\geq 0}$, and $g^{(0)} = \mathbb{I}$.

2. Problem Statement

Consider the system

$$x^+ = f(x, u), \tag{1}$$

with state $x \in \mathbb{R}^n$, control input $u \in \mathcal{U}(x)$ where $\mathcal{U}(x) \subseteq \mathbb{R}^m$ is the set of admissible inputs, and $f : \mathcal{W} \to \mathbb{R}^n$ where $\mathcal{W} := \{(x, u) : x \in \mathbb{R}^n, u \in \mathcal{U}(x)\}$. We use $\phi(k, x, \boldsymbol{u}|_k)$ to denote the solution to system (1) at time $k \in \mathbb{Z}_{\geq 0}$ with initial condition x and inputs sequence $\boldsymbol{u}|_k = [u_0, u_1, \ldots, u_{k-1}]$, with the convention $\phi(0, x, \boldsymbol{u}|_0) = x$.

We consider the infinite-horizon cost

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

where $x \in \mathbb{R}^n$ is the initial state, $\boldsymbol{u}$ is an infinite sequence of admissible inputs, $\ell : \mathcal{W} \to \mathbb{R}_{\geq 0}$ is the stage cost. Finding an infinite sequence of inputs which minimizes (2) given $x \in \mathbb{R}^n$ is very difficult in general. Therefore, we instead generate sequences of admissible inputs that *nearly* minimize (2), in a sense made precise below, while ensuring the stability of the closed-loop system. For this purpose, we consider VI, see e.g. (Bertsekas, 2012). VI is an iterative procedure based on Bellman equation, which we briefly recall next. Assuming the optimal value function, denoted V_∞, exists for any $x \in \mathbb{R}^n$, the Bellman equation is

$$V_\infty(x) = \min_{u \in \mathcal{U}(x)} \Big\{ \ell(x, u) + V_\infty(f(x, u)) \Big\}. \tag{3}$$

If we could solve (3) and find V_∞, it would then be easy to derive an optimal policy, by computing the $\arg\min$ corresponding to the right hand-side of (3). However, it is in general very difficult to

solve (3). VI provides an iterative procedure based on (3) instead, which allows obtaining value functions (and associated control inputs), which converge to V_∞. Hence, given an initial cost function $V_{-1} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, VI generates a sequence of value functions V_d, $d \in \mathbb{Z}_{\geq 0}$, for any $x \in \mathbb{R}^n$, by iterating

$$V_d(x) := \min_{u \in \mathcal{U}(x)} \Big\{ \ell(x,u) + V_{d-1}(f(x,u)) \Big\}. \tag{4}$$

For any $d \in \mathbb{Z}_{\geq 0}$, the associated input, also called policy, is defined as, for any $x \in \mathbb{R}^n$,

$$u_d^*(x) \in \arg\min_{u \in \mathcal{U}(x)} \Big\{ \ell(x,u) + V_{d-1}(f(x,u)) \Big\}, \tag{5}$$

which may be set-valued. The convergence of V_d, $d \in \mathbb{Z}_{\geq 0}$, to V_∞ in (3) is ensured under mild conditions in (Bertsekas, 2017). In the sequel we make assumptions that ensure that the $\arg\min$ in (5) exists for each $x \in \mathbb{R}^n$.

In practice, we often stop iterating VI when a stopping criterion is verified, such as, for instance, when for any $x \in \mathbb{R}^n$,

$$V_d(x) - V_{d-1}(x) \leq \varepsilon, \tag{6}$$

where $\varepsilon \in \mathbb{R}_{>0}$, see, e.g., (Bertsekas, 2012; Sutton and Barto, 2017; Pang et al., 2019; Kiumarsi et al., 2017). However, this stopping criterion leaves much to be desired in control applications, for the following reasons: (i) it is not yet established how ε impacts the stability properties of the closed-loop system; (ii) tools to bound the mismatch between V_d and V_∞ for this stopping criterion often requires a discount factor in cost function (2), which impacts stability, as shown in (Postoyan et al., 2017, 2019); (iii) when V_d is radially unbounded, i.e. $V_d(x) \to \infty$ when $|x| \to \infty$, this stopping criterion is in general impossible to verify for all $x \in \mathbb{R}^n$. When the system is linear and the cost quadratic, as in (Arnold and Laub, 1984; Anderson and Moore, 2007; Jiang and Jiang, 2012; Bian and Jiang, 2016), the convergence to the optimal cost function is shown to be quadratic and often the stopping criterion is instead of the form $V_d(x) - V_{d-1}(x) \leq |\varepsilon||x|^2$. However, the link between the value of ε and resulting near-optimality and stability guarantees is not established, and in practice it is implicitly assumed that parameter ε is small enough.

We consider VI terminated by a general stopping criterion. That is, for any $x \in \mathbb{R}^n$,

$$V_d(x) - V_{d-1}(x) \leq c_{\text{stop}}(\varepsilon, x), \tag{7}$$

where $c_{\text{stop}}(\varepsilon, x) \geq 0$ is a stopping function, which we design and which may depend on state vector x and a vector of tuneable parameters $\varepsilon \in \mathbb{R}^{n_\varepsilon}$ with $n_\varepsilon \in \mathbb{Z}_{>0}$. The design of c_{stop} is explained in Section 3. In that way, we cover the above examples as particular cases, namely $c_{\text{stop}}(x, \varepsilon) = |\varepsilon|$ and $c_{\text{stop}}(\varepsilon, x) = |\varepsilon||x|^2$ and allow considering more general ones, e.g. $c_{\text{stop}}(\varepsilon, x) = \max\{|\varepsilon_1|, |\varepsilon_2||x|^2\}$ where $(\varepsilon_1, \varepsilon_2) := \varepsilon \in \mathbb{R}^2$ or $c_{\text{stop}}(\varepsilon, x) = x^\top S(\varepsilon) x$ for some positive definite matrix $S(\varepsilon)$ with $\varepsilon \in \mathbb{R}^{n_\varepsilon}$ and $n_\varepsilon \in \mathbb{Z}_{>0}$. The main novelty of this work is the provided explicit link between $c_{\text{stop}}(\varepsilon, x)$, near-optimality and stability guarantees. As a result, we can tune ε for the desired near-optimality and stability properties, and the algorithm stops when the cost (hence, the generated inputs) are such that these properties are verified.

The analysis relies on the next assumption[1] like in e.g., (Grimm et al., 2005; Postoyan et al., 2017; Granzotto et al., 2020a).

1. The assumption is stated globally, for any $x \in \mathbb{R}^n$ and $u \in \mathcal{U}(x)$. We leave for future work the case where the assumption holds on compact sets.

Standing Assumption 1 (SA1) *There exist $\alpha_V, \alpha_W \in \mathcal{K}_\infty$ and continuous function $\sigma : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ such that the following conditions hold.*

(i) For any $x \in \mathbb{R}^n$, there exists an infinite sequence of admissible inputs $\boldsymbol{u}^*_\infty(x)$, called optimal input sequence, which minimizes (2), i.e. $V_\infty(x) = J_\infty(x, \boldsymbol{u}^*_\infty(x))$, and $V_\infty(x) \leq \alpha_V(\sigma(x))$.

(ii) For any $(x, u) \in \mathcal{W}$, $\alpha_W(\sigma(x)) \leq \ell(x, u)$. □

Function $\sigma : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ in SA1 is a "measuring" function that we use to define stability, which depends on the problem. For instance, by defining $\sigma = |\cdot|, \sigma = |\cdot|^2$ or $\sigma : x \mapsto x^\top Q x$ with $Q = Q^\top > 0$, one would be studying the stability of the origin, and by taking $\sigma = |\cdot|_{\mathcal{A}}$, one would study stability of non-empty compact set $\mathcal{A} \subset \mathbb{R}^n$. General conditions to ensure the first part of item (i), i.e. the fact that $V_\infty(x)$ is finite for any $x \in \mathbb{R}^n$ and the existence of optimal inputs, can be found in (Keerthi and Gilbert, 1985). The second part of item (i) is related to the stabilizability of system (1) with respect to stage cost ℓ in relation to σ. Indeed, it is shown in (Grimm et al., 2005, Lemma 1) that, for instance, when the stage cost $\ell(x, u)$ is uniformly globally exponentially controllable to zero with respect to σ for system (1), see (Grimm et al., 2005, Definition 2), then item (i) of SA1 is satisfied. Hence, item (i) of SA1 is ensured when we know a stabilizing, but not necessarily optimal, input sequence that makes ℓ exponentially decrease along solutions of (1). We do not need to know V_∞ to guarantee the last inequality in item (i) of SA1. Indeed, it suffices to find, for any $x \in \mathbb{R}^n$, a sequence of inputs $\boldsymbol{u}(x)$, such the associated infinite-horizon costs verifies $J(x, \boldsymbol{u}(x)) \leq \alpha_V(\sigma(x))$ for some $\alpha_V \in \mathcal{K}_\infty$. Then, since V_∞ is the optimal value function, for any $x \in \mathbb{R}^n$, $V_\infty(x) \leq J(x, \boldsymbol{u}(x)) \leq \alpha_V(\sigma(x))$. On the other hand, item (ii) of SA1 is a detectability property of the stage cost ℓ with respect to σ, as when $\ell(x, u)$ is small, so is $\sigma(x)$.

We are ready to explain how to design the stopping criterion in (7).

3. Stopping criterion design

3.1. Key observation

We start with the known observation (Granzotto, 2019; Bertsekas, 2005) that, given[2] $V_{-1} = 0$, at each iteration $d \in \mathbb{Z}_{\geq 0}$, VI generates the optimal value function for the finite-horizon cost

$$J_d(x, \boldsymbol{u}_d) := \sum_{k=0}^{d} \ell(\phi(k, x, \boldsymbol{u}_d|_k), u_k), \tag{8}$$

where $\boldsymbol{u}_d = [u_0, u_1, ..., u_d]$ are admissible inputs. We assume below that the minimum of (8) exists with relation to $\boldsymbol{u}_d$ for any $x \in \mathbb{R}^n$ and $d \in \mathbb{Z}_{\geq 0}$.

Standing Assumption 2 (SA2) *For every $d \in \mathbb{Z}_{\geq 0}$, $x \in \mathbb{R}^n$, there exists $\boldsymbol{u}^\star_d(x)$ such that $V_d(x) = J_d(x, \boldsymbol{u}^\star_d) = \min_{\boldsymbol{u}_d} J_d(x, \boldsymbol{u}_d)$.* □

SA2 is for instance verified when f and ℓ are continuous and $\mathcal{U}(x) = \mathcal{U}$ is a compact set. More general conditions to verify SA2 can be found in e.g. (Keerthi and Gilbert, 1985). For the sake of convenience, we employ the following notation for the technical aspects of this paper. For any $k \in \{0, 1, \ldots, d\}$ and $x \in \mathbb{R}^n$, we denote $\ell^*_d(k, x) := \ell(\phi(k, x, \boldsymbol{u}^\star_d(x)|_k), u_k)$, where $\phi(k, x, \boldsymbol{u}^\star_d(x)|_k)$ is the solution to system (1) with optimal inputs for cost $V_d(x)$, so that $V_d(x) = \sum_{k=0}^{d} \ell^*_d(k, x)$.

Before we explain how to design c_{stop}, we state the next property which plays a key role in the forthcoming analysis.

2. The case where $V_{-1} \neq 0$ will be investigated in further work.

Proposition 1 *For any $x \in \mathbb{R}^n$ and $d \in \mathbb{Z}_{\geq 0}$, $\ell_d^*(d,x) \leq V_d(x) - V_{d-1}(x)$.* □

When the stopping criterion (7) is verified, i.e. $V_d(x) - V_{d-1}(x) \leq c_{\text{stop}}(\varepsilon, x)$, then $\ell_d^*(d,x) \leq c_{\text{stop}}(\varepsilon, x)$ in view of Proposition 1. Therefore, $c_{\text{stop}}(\varepsilon, x)$ is an upper-bound on the value of stage cost $\ell_d^*(d,x)$. By item (ii) of SA1, this implies that we also have an upper-bound for *d-horizon* state measure $\sigma(\phi(d, x, \boldsymbol{u}_d^*(x)|_d))$, namely $\sigma(\phi(d, x, \boldsymbol{u}_d^*(x)|_d)) \leq \alpha_W^{-1}(c_{\text{stop}}(\varepsilon, x))$, which can be made as small as desired by reducing $c_{\text{stop}}(\varepsilon, x)$, which, again, we design. We exploit this property to analyse the near-optimality and the stability of the closed-loop system. Having said that, the challenges are: (i) to show that condition (7) is indeed verified for any $x \in \mathbb{R}^n$ and some $d \in \mathbb{Z}_{\geq 0}$; (ii) to select c_{stop} to ensure stability properties when closing the loop of system (1) with inputs (5); (iii) to study the impact of c_{stop} on the performance, that is, the cost along solutions, of the closed-loop system.

3.2. Satisfaction of the stopping criterion

We design c_{stop} to satisfy the next assumption; suitable examples are given afterwards.

Assumption 1 *Out of the two next properties, one holds.*

(i) For any $\varepsilon \in \mathbb{R}^{n_\varepsilon}$, there is $\underline{\epsilon} > 0$ such that, for any $x \in \mathbb{R}^n$, $c_{stop}(\varepsilon, x) \geq \underline{\epsilon}$.

(ii) There exist $L, a_V, a_W > 0$, such that SA1 holds with $\alpha_V(s) \leq a_V s$ and $\alpha_W(s) \geq a_W s$ for any $s \in [0, L]$. Furthermore, for any $\varepsilon \in \mathbb{R}^{n_\varepsilon}$, there is $\underline{\epsilon} > 0$ such that for any $x \in \mathbb{R}^n$, $c_{stop}(\varepsilon, x) \geq \underline{\epsilon}\sigma(x)$. □

Item (i) of Assumption 1 can be ensured by taking $c_{\text{stop}}(\varepsilon, x) = |\varepsilon| + \tilde{c}_{\text{stop}}(x, \varepsilon)$ with $\tilde{c}_{\text{stop}}(x, \varepsilon) \geq 0$ for any $x \in \mathbb{R}^n$, $\varepsilon \in \mathbb{R}^{n_\varepsilon}$, which covers (6), to give an example. Item (ii) of Assumption 1 means that the functions α_V, α_W in SA1 can be upper-bounded, respectively lower-bounded, by linear functions on the interval $[0, L]$. These conditions allow to select c_{stop} such that $c_{\text{stop}}(\varepsilon, x) \to 0$ when $\sigma(x) \to 0$ with $x \in \mathbb{R}^n$, contrary to item (i) of Assumption 1, that is, c_{stop} may vanish on set $\{x : \sigma(x) = 0\}$. This is important to provide stronger stability and performance properties for systems whose inputs are given by our VI scheme as shown in Section 4. Under item (ii) of Assumption 1, we can design c_{stop} as, e.g., $c_{\text{stop}}(\varepsilon, x) = |\varepsilon|\sigma(x)$, $c_{\text{stop}}(\varepsilon, x) = \min\{|\varepsilon_1|, |\varepsilon_2||x|^2\}$ where $(\varepsilon_1, \varepsilon_2) =: \varepsilon \in \mathbb{R}^2$ or $c_{\text{stop}}(\varepsilon, x) = x^\top S(\varepsilon) x$ for some positive definite matrix $S(\varepsilon)$ as mentioned before.

The next theorem ensures the existence of $d \in \mathbb{Z}_{\geq 0}$ such that, for any $x \in \mathbb{R}^n$, (7) holds based on Assumption 1.

Theorem 2 *Suppose Assumption 1 holds. Then, for any $\Delta > 0$ there exists $d \in \mathbb{Z}_{\geq 0}$ such that, for any $x \in \{z \in \mathbb{R}^n : \sigma(z) \leq \Delta\}$, (7) holds. Moreover, when item (ii) of Assumption 1 holds with $L = \infty$, there exists $d \in \mathbb{Z}_{\geq 0}$ such that, for any $x \in \mathbb{R}^n$, (7) is satisfied.* □

Theorem 2 guarantees the stopping condition in (7) is always satisfied by iterating the VI algorithm sufficiently many times, and that the required number of iterations is uniform over sets of initial conditions of the form $\{x : \sigma(x) \leq \Delta\}$ for given $\Delta > 0$ in general, unless item (ii) of Assumption 1 holds with $L = \infty$, in which case there exists a common, global, d for any $x \in \mathbb{R}^n$. Note that, while the proof of Theorem 2 provides a conservative estimate of d such that (7) is verified, see (Granzotto et al., 2020c), this horizon estimate is not utilized in the stopping criterion, which in turn implies that VI stops with smaller horizon, in general, as illustrated in Section 5.

In the following, we denote the cost calculated at iteration d as $V_\varepsilon(x) := V_d(x)$, like in (Granzotto et al., 2020b), to emphasize that the cost returned is parameterized by ε via $c_{\text{stop}}(\varepsilon, \cdot)$, and denote by $\boldsymbol{u}^\star_\varepsilon(x)$ an associated optimal sequence of inputs, i.e.

$$V_\varepsilon(x) = J_d(x, \boldsymbol{u}^\star_\varepsilon(x)). \tag{9}$$

We are ready to state the main results.

4. Main results

In this section, we analyze the near-optimality properties of VI with the stopping criterion in (7). We then provide conditions under which system (1), whose inputs are generated by applying the state-feedback $\boldsymbol{u}^\star_\varepsilon(x)$ in receding-horizon fashion, exhibits stability properties. Afterwards, the cost function (8) along the solutions of the induced closed-loop system are analysed, which we refer to by performance or running-cost (Grüne and Rantzer, 2008).

4.1. Relationship between V_ε and V_∞

A key question is how far is V_ε from V_∞ when we stop VI using (7). Since $\ell(x, u)$ is not constrained to take values in a given compact set, and we do not consider discounted costs, the tools found in the dynamic programming literature to analyze this relationship are no longer applicable, see (Bertsekas, 2012). We overcome this issue by exploiting SA1, and adapting the results of (Granzotto et al., 2020a) with the stopping criterion and Proposition 1 in the next theorem.

Theorem 3 *Suppose Assumption 1 holds. For any $\varepsilon \in \mathbb{R}^{n_\varepsilon}$, $\Delta > 0$ and $x \in \{z \in \mathbb{R}^n, \sigma(z) \leq \Delta\}$,*

$$V_\varepsilon(x) \leq V_\infty(x) \leq V_\varepsilon(x) + v_\varepsilon(x), \tag{10}$$

where $v_\varepsilon(x) := \alpha_V \circ \alpha_W^{-1}(c_{stop}(\varepsilon, x))$ with α_V, α_W from SA1. Moreover, when item (ii) of Assumption holds with $L = \infty$, we accept $\Delta = \infty$ and $v_\varepsilon(x) \leq \frac{a_V}{a_W} c_{stop}(\varepsilon, x)$. □

The lower-bound in (10) trivially holds from the optimality of $V_\varepsilon(x) = V_d(x)$ for some $d < \infty$, and the fact that $\ell(x, u) \geq 0$ for any $x \in \mathbb{R}^n$ and $u \in \mathcal{U}(x)$. The upper-bound, on the other hand, implies that the infinite-horizon cost is at most $v_\varepsilon(x)$ away from the finite-horizon $V_\varepsilon(x)$. The error term $v_\varepsilon(x)$ is small when $c_{\text{stop}}(\varepsilon, x)$ is small as $\alpha_V \circ \alpha_W^{-1} \in \mathcal{K}_\infty$. Given that we know α_V, α_W^{-1} a priori, and we are free to design c_{stop} as wanted, we can therefore directly make $V_\varepsilon(x)$ as close as desired to $V_\infty(x)$ by adjusting c_{stop}; the price to pay will be more computations. Moreover, when item (ii) of Assumption holds with $L = \infty$, inequality (10) is verified for every $x \in \mathbb{R}^n$.

4.2. Stability

We now consider the scenario where system (1) is controlled in a receding-horizon fashion by inputs that calculate cost (9). That is, at each time instant $k \in \mathbb{Z}_{\geq 0}$, the first element of optimal sequence $\boldsymbol{u}^\star_\varepsilon(x_k)$, calculated by VI, is then applied to system (1). This leads to the closed-loop system

$$x^+ \in f(x, \mathcal{U}^*_\varepsilon(x)) =: F^*_\varepsilon(x), \tag{11}$$

where $f(x, \mathcal{U}^*_\varepsilon(x))$ is the set $\{f(x, u) : u \in \mathcal{U}^*_\varepsilon(x)\}$ and $\mathcal{U}^*_\varepsilon(x) := \big\{u_0 : \exists u_1, \ldots, u_d \in \mathcal{U}(x)$ such that $V_\varepsilon(x) = J_d(x, [u_0, \ldots, u_d])\big\}$ is the set of the first input of d-horizon optimal input sequences at x, with d as defined in (7). We denote by $\phi(k, x)$ a solution to (11) at time $k \in \mathbb{Z}_{\geq 0}$ with initial condition $x \in \mathbb{R}^n$, with some abuse of notation.

We assume next that c_{stop} can be made as small as desirable by taking $|\varepsilon|$ sufficiently small. As we are free to design c_{stop} as wanted, this is without loss of generality.

Assumption 2 *There exists $\theta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, with $\theta(\cdot, s) \in \mathcal{K}$ and $\theta(s, \cdot)$ non-decreasing for any $s > 0$, such that $c_{stop}(\varepsilon, x) \leq \theta(|\varepsilon|, \sigma(x))$ for any $x \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}^{n_\varepsilon}$.* □

Example of functions c_{stop} which satisfy Assumption 2 are $c_{\text{stop}}(\varepsilon, x) = |\varepsilon|\sigma(x)$, $c_{\text{stop}}(\varepsilon, x) = \max\{|\varepsilon_1|\alpha(\sigma(x)), |\varepsilon_2|\}$ for $\varepsilon = (\varepsilon_1, \varepsilon_2) \in \mathbb{R}^2$, $\alpha \in \mathcal{K}$ and $x \in \mathbb{R}^n$ to give a few.

The next theorem provides stability guarantees for system (11).

Theorem 4 *Consider system (11) and suppose c_{stop} verifies Assumptions 1 and 2. There exists $\beta \in \mathcal{KL}$ such that, for any $\delta, \Delta > 0$, there exists $\varepsilon^* > 0$ such that for any $x \in \{z \in \mathbb{R}^n : \sigma(z) \leq \Delta\}$ and $\varepsilon \in \mathbb{R}^{n_\varepsilon}$ with $|\varepsilon| < \varepsilon^*$, any solution $\phi(\cdot, x)$ to system (11) satisfies, for all $k \in \mathbb{Z}_{\geq 0}$, $\sigma(\phi(k, x)) \leq \max\{\beta(\sigma(x), k), \delta\}$.* □

Theorem 4 provides a uniform semiglobal practical stability property for the set $\{z : \sigma(z) = 0\}$. This implies that solutions to (11), with initial state x such that $\sigma(x) \leq \Delta$, where Δ is any given (arbitrarily large) strictly positive constant, will converge to the set $\{z : \sigma(z) \leq \delta\}$, where δ is any given (arbitrarily small) strictly positive constant, by taking ε^* sufficiently close to 0, thereby making c_{stop} sufficiently small. An explicit formula for ε^* is given in the proof of Theorem 4 in (Granzotto et al., 2020c), which is nevertheless subject to some conservatism. The result should rather be appreciated qualitatively, in the sense that Theorem 4 holds for small enough ε^*.

Under stronger assumptions, global exponential stability is ensured as shown in the next corollary.

Corollary 5 *Suppose item (ii) of Assumption 1 holds and that $c_{stop}(\varepsilon, x) \leq |\varepsilon|\sigma(x)$ for any $x \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}^{n_\varepsilon}$. Let $\varepsilon^* > 0$ be such that $\varepsilon^* < \frac{a_W^2}{a_V}$. Then, for any $x \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}^{n_\varepsilon}$ such that $|\varepsilon| \leq \varepsilon^*$, any solution $\phi(\cdot, x)$ to system (11) satisfies $\sigma(\phi(k, x)) \leq \frac{a_V}{a_W}\left(1 - \frac{a_W^2 - |\varepsilon| a_V}{a_V a_W}\right)^k \sigma(x)$ for all $k \in \mathbb{Z}_{\geq 0}$.* □

Corollary 5 ensures a uniform global exponential stability property of set $\{x : \sigma(x) = 0\}$ for system (11). Indeed, in Corollary 5, the decay rate is given by $1 - \frac{a_W^2 - |\varepsilon| a_V}{a_V a_W}$ and take values in $(0, 1)$ as $|\varepsilon| \leq \varepsilon^* < \frac{a_W^2}{a_V}$ as required by Corollary 5, hence $\left(1 - \frac{a_W^2 - |\varepsilon|}{a_V a_W}\right)^k \to 0$ as $k \to \infty$. Furthermore, the estimated decay rate can be tuned via ε from 1 to $1 - \frac{a_W}{a_V}$ as $|\varepsilon|$ decreases to zero. We can therefore make the decay smaller by adjusting c_{stop}, as in Theorem 3. Hence, by tuning ε, we can tune how fast the closed-loop converges to the attractor $\{x : \sigma(x) = 0\}$, and the price to pay is more computations in general.

4.3. Policy performance guarantees

In Section 4.1, we have provided relationships between the finite-horizon cost V_ε and the infinite-horizon cost V_∞. This is an important feature of VI, but this does not directly provide us with information on the actual value of the cost function (2) along solutions to (11). Therefore, we analyse the running cost (Grüne and Rantzer, 2008) defined as

$$\mathcal{V}_\varepsilon^{\text{run}}(x) := \left\{ \sum_{k=0}^{\infty} \ell_{\mathcal{U}_\varepsilon^*(\phi(k,x))}(\phi(k, x)) : \phi(\cdot, x) \text{ is a solution to (11)} \right\}, \tag{12}$$

where $\ell_{\mathcal{U}_\varepsilon^*(\phi(k,x))}(\phi(k, x))$ is the actual stage cost incurred at time step k. It has to be noted that $\mathcal{V}_\varepsilon^{\text{run}}(x)$ is a set, since solutions of (11) are not necessarily unique. Each element $V_\varepsilon^{\text{run}}(x) \in \mathcal{V}_\varepsilon^{\text{run}}(x)$

corresponds then to the cost of a solution of (11). Clearly, $V_\varepsilon^{\text{run}}(x)$ is not necessarily bounded, as the stage costs may not decrease to 0 in view of Theorem 4. Indeed, only practical convergence is ensured in Theorem 4 in general. On the other hand, when the set $\{x \in \mathbb{R}^n : \sigma(x) = 0\}$ is globally exponentially stable as in Corollary 5, the elements of $\mathcal{V}_\varepsilon^{\text{run}}(x)$ in (12) are bounded and satisfy the next property.

Theorem 6 *Consider system (11) and suppose the conditions of Corollary 5 hold. For any ε such that $|\varepsilon| < \varepsilon^*$, $x \in \mathbb{R}^n$, and $V_\varepsilon^{run}(x) \in \mathcal{V}_\varepsilon^{run}(x)$, it follows that $V_\infty(x) \leq V_\varepsilon^{run}(x) \leq V_\infty(x) + w_\varepsilon \sigma(x)$, with $w_\varepsilon := \dfrac{a_V^3}{a_W} \dfrac{|\varepsilon|}{a_W^2 - a_V|\varepsilon|}$, where the constants come from Corollary 5.* □

The inequality $V_\infty(x) \leq V_\varepsilon^{\text{run}}(x)$ of Theorem 6 directly follows from the optimality of V_∞. The inequality $V_\varepsilon^{\text{run}}(x) \leq V_\infty(x) + w_\varepsilon \sigma(x)$ provides a relationship between the running cost $V_\varepsilon^{\text{run}}(x)$ and the infinite-horizon cost at state x, $V_\infty(x)$. The inequality $V_\varepsilon^{\text{run}}(x) \leq V_\infty(x) + w_\varepsilon \sigma(x)$ confirms the intuition coming from Theorem 3 that a smaller stopping criterion leads to tighter near-optimality guarantees. That is, when $|\varepsilon| \to 0$, $w_\varepsilon \to 0$ and $V_\varepsilon^{\text{run}}(x) \to V_\infty(x)$ for any $x \in \mathbb{R}^n$, provided that Corollary 5 holds. However, this comes at the price of more iterations and thus more computations to satisfy a tighter stopping criterion in (7). In contrast with Theorem 3, stability of system (11) is essential in Theorem 6. Indeed, the term $\frac{1}{a_W^2 - a_V|\varepsilon|}$ in the expression of w_ε shows that the running cost is large when $|\varepsilon|$ is close to a_W^2/a_V, hence, when stability is not guaranteed, the running cost might be unbounded.

5. Illustrative Example

We consider the discrete cubic integrator, also seen in (Grimm et al., 2005; Granzotto et al., 2020a), which is given by $(x_1^+, x_2^+) = (x_1 + u, x_2 + u^3)$ where $(x_1, x_2) := x \in \mathbb{R}^2$ and $u \in \mathbb{R}$. Let $\sigma(x) = |x_1|^3 + |x_2|$ and consider cost (8) with $\ell(x, u) = |x_1|^3 + |x_2| + |u|^3$ for any $(x, u) \in \mathbb{R}^2 \times \mathbb{R}$. It is shown in (Granzotto et al., 2020a) that SA1 holds with $\alpha_V = 14\mathbb{I}$ and $\alpha_W := \mathbb{I}$.

Because it is notoriously difficult to exactly compute $V_d(x)$ and associated sequence of optimal inputs for every $x \in \mathbb{R}^2$, we use an approximate scheme. In particular, we rely on a simple finite difference approximation, with $N = 340^2$ points equally distributed in $[-10, 10] \times [-10^3, 10^3]$ for the state space or, equivalently, $\{x \in \mathbb{R}^n : \sigma(x) \leq 2000\}$, and 909 equally distributed quantized inputs in $[-20, 20]$ centered at 0. We consider three types of stopping criteria for which ε is a scalar. For each stopping criterion, we discuss the type of guaranteed stability and we provide in Table 1 the corresponding horizon for different values of ε, which is related to the computation cost. Then, for each horizon, we give in Table 2 estimates of the running cost for initial condition $x = (10, -10^3)$, by computing the sum in (12) up to $k = 40$ instead of $k = \infty$, as well as the value of $\sigma(\phi(40, x))$ to evaluate the convergence accuracy of the corresponding policy.

We first take the uniform stopping criterion uniform stopping criterion as in (6), like in, e.g.,(Bertsekas, 2012; Sutton and Barto, 2017; Pang et al., 2019; Kiumarsi et al., 2017; Liu et al., 2015), i.e. $c_{\text{stop}}(\varepsilon, x) := |\varepsilon|$, with different values of ε. In this case, we have no global exponential stability or performance guarantees like in Corollary 5 and Theorem 6 a priori. Only near-optimal guarantees as in Theorem 3 and semiglobal practical stability as in Theorem 4 hold. For instance, by taking $\varepsilon = 0.01$, Theorem 3 holds with $v_\varepsilon(x) = 14 \cdot 0.01 = 0.14$ for any $x \in \mathbb{R}^n$.

We also consider the following relative stopping criterion, for any $x \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}$, $c_{\text{stop}}(\varepsilon, x) := |\varepsilon|\sigma(x)$. The exponential stability of Corollary 5 holds for any $\varepsilon \in \mathbb{R}$ such that $|\varepsilon| < \frac{a_W^2}{a_V} = \frac{1}{14}$ in this

		ε						
		10	0.75	0.1	0.075	0.05	0.025	0.005
$c_{\text{stop}}(\varepsilon, x)$:	$\lvert\varepsilon\rvert$	$d=6$	$d=7$	$d=8$	$d=8$	$d=8$	$d=8$	$d=9$
	$\lvert\varepsilon\rvert\sigma(x)$	$d=0$	$d=1$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$
	$\lvert\varepsilon\rvert \min\{\sigma(x), 1\}$	$d=6$	$d=7$	$d=8$	$d=8$	$d=8$	$d=8$	$d=9$

Table 1: Required iterations to fulfill each stopping criteria for $N = 340^2$ points equally distributed in $\{z \in \mathbb{R}^n : \sigma(z) \leq 2000\}$.

	d								
	0	1	3	4	5	6	7	8	9
$V_d^{\text{run}}(x)$	77313	45497	19931	19965	19802	20090	20359	20261	20261
$\sigma(\phi(40, x))$	1982	1138	2.56	2.84	1.71	2.25	1.84	1.62	1.62

Table 2: Estimation of the running cost $V_d^{\text{run}}(x)$ for $x = (10, -10^3)$ and the value of $\sigma(\phi(40, x))$.

case. Moreover, we have near-optimality and performance properties as in Theorems 3 and 6, which were not available for the previous stopping criterion $c_{\text{stop}}(\varepsilon, x) = |\varepsilon|$. Moreover, for $\varepsilon = 0.01 < \frac{1}{14}$, Theorem 3 holds with $v_\varepsilon(x) = 14 \cdot 0.01 \cdot \sigma(x) = 0.14\sigma(x)$ for any $x \in \mathbb{R}^n$, which is small when $\sigma(x)$ is small, and vice versa. Compared to the previous stopping criterion, which leads to constant guaranteed near-optimality bound, here we have better guarantees when $\sigma(x)$ is small (and worse ones when $\sigma(x)$ is large). We observe less computations for better a priori near-optimality properties for states near the attractor, i.e. when $\sigma(x) < 1$, when compared to the previous stopping criterion. We finally consider the mixed stopping criterion $c_{\text{stop}}(\varepsilon, x) := |\varepsilon| \min\{\sigma(x), 1\}$, which provides better near-optimality guarantees than both considered stopping criteria. We see from Table 1, and Table 2, that by increasing iterations, we usually obtain smaller and thus better running costs as well as tighter convergence properties.

Compared to previous work (Granzotto et al., 2020a), where stability properties to (approximate) value iteration are given, we require a smaller number of iterations. Indeed, in view of (Granzotto et al., 2020a, Corollary 2), $d \geq \bar{d} = \left\lfloor \frac{0 - \ln 14^2}{\ln 13 - \ln 14} \right\rfloor = 71$. Of course, this analysis is conservative and a different derivation of α_V might provide different bounds on $\bar{d}$. Here, as the algorithm is free to choose the required number of iterations via the stopping criterion, we significantly reduce its conservatism. This induces smaller computational complexity, as, e.g. for $\varepsilon = 0.01 < \frac{1}{14}$, exponential stability is ensured with the stoping criterion verified at $d = 6$, that is, 8.5% of iterations required by the lower bound $\bar{d} = 71$ of (Granzotto et al., 2020a).

6. Concluding remarks

Future work includes relaxing the initial condition for VI and the main assumptions. Another direction is extending the work towards stochastic problems and online algorithms, towards the final goal of stability-based computational-performance tradeoffs in reinforcement learning.

References

B. D. O. Anderson and J. B. Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

W. F. Arnold and A. J. Laub. Generalized eigenproblem algorithms and software for algebraic riccati equations. *Proceedings of the IEEE*, 72(12):1746–1754, 1984.

F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, pages 908–918, 2017.

D. P. Bertsekas. Dynamic programming and suboptimal control: A survey from ADP to MPC. *European Journal of Control*, 11(4-5):310–334, 2005.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Nashua, USA, 4th edition, 2012.

D. P. Bertsekas. Value and policy iterations in optimal control and adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):500–509, 2017. doi: 10.1109/TNNLS.2015.2503980.

T. Bian and Z.-P. Jiang. Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. *Automatica*, 71:348 – 360, 2016. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2016.05.003.

L. Buşoniu, T. de Bruin, D. Tolić, J. Kober, and I. Palunko. Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control*, 46:8 – 28, 2018. ISSN 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2018.09.005.

M. Granzotto. *Near-optimal control of discrete-time nonlinear systems with stability guarantees*. PhD thesis, Université de Lorraine, 2019. URL http://www.theses.fr/2019LORR0301.

M. Granzotto, R. Postoyan, L. Buşoniu, D. Nešić, and J. Daafouz. Finite-horizon discounted optimal control: stability and performance. *IEEE Transactions on Automatic Control*, 2020a. doi: 10.1109/TAC2020.2985904.

M. Granzotto, R. Postoyan, L. Buşoniu, D. Nešić, and J. Daafouz. Stable near-optimal control of nonlinear switched discrete-time systems: a planning-based approach. In *Submitted to journal publication*, 2020b.

M. Granzotto, R. Postoyan, D. Nešić, L. Buşoniu, and J. Daafouz. When to stop value iteration: stability and near-optimality versus computation. In *Submitted to arXiv*, 2020c.

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. ISSN 0018-9286. doi: 10.1109/TAC.2005.847055.

L. Grüne and A. Rantzer. On the infinite horizon performance of receding horizon controllers. *IEEE Transactions on Automatic Control*, 53(9):2100–2111, 2008. ISSN 0018-9286. doi: 10.1109/TAC.2008.927799.

A. Heydari. Revisiting approximate dynamic programming and its convergence. *IEEE Transactions on Cybernetics*, 44(12):2733–2743, 2014. doi: 10.1109/TCYB.2014.2314612.

A. Heydari. Analysis of stabilizing value iteration for adaptive optimal control. In *2016 American Control Conference (ACC)*, pages 5746–5751, 2016. doi: 10.1109/ACC.2016.7526570.

A. Heydari. Stability analysis of optimal adaptive control under value iteration using a stabilizing initial policy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9):4522–4527, 2017.

A. Heydari. Stability analysis of optimal adaptive control using value iteration with approximation errors. *IEEE Transactions on Automatic Control*, 2018. ISSN 0018-9286. doi: 10.1109/TAC.2018.2790260.

J.-F. Hren and R. Munos. Optimistic planning of deterministic systems. In *European Workshop on Reinforcement Learning*, pages 151–164, Villeneuve d'Ascq, France, 2008.

Y. Jiang and Z.-P. Jiang. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 48(10):2699 – 2704, 2012. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2012.06.096.

S. Keerthi and E. Gilbert. An existence theorem for discrete-time infinite-horizon optimal control problems. *IEEE Transactions on Automatic Control*, 30(9):907–909, 1985. ISSN 0018-9286. doi: 10.1109/TAC.1985.1104084.

B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang. H∞ control of linear discrete-time systems: Off-policy reinforcement learning. *Automatica*, 78:144–152, 2017.

F. L. Lewis and D. Vrabie. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 9(3):32–50, 2009. doi: 10.1109/MCAS.2009.933854.

D. Liu, H. Li, and D. Wang. Error bounds of adaptive dynamic programming algorithms for solving undiscounted optimal control problems. *IEEE Transactions on Neural Networks and Learning Systems*, 26(6):1323–1334, 2015. doi: 10.1109/TNNLS.2015.2402203.

B. Pang, T. Bian, and Z.-P. Jiang. Adaptive dynamic programming for finite-horizon optimal control of linear time-varying discrete-time systems. *Control Theory and Technology*, 17(1):73–84, 2019.

A. Pavlov, I. Shames, and C. Manzie. Early termination of NMPC interior point solvers: Relating the duality gap to stability. In *2019 18th European Control Conference (ECC)*, pages 805–810, 2019. doi: 10.23919/ECC.2019.8795629.

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. ISSN 0018-9286. doi: 10.1109/TAC.2016.2616644.

R. Postoyan, M. Granzotto, L. Buşoniu, B. Scherrer, D. Nešić, and J. Daafouz. Stability guarantees for nonlinear discrete-time systems controlled by approximate value iteration. In *IEEE Conference on Decision and Control*, Nice, France, 2019.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, USA, 2nd edition, 2017.

Q. Wei, D. Liu, and H. Lin. Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems. *IEEE Transactions on Cybernetics*, 46(3):840–853, 2015.