Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

A reinforcement learning method with closed-loop stability guarantee

Pavel Osinenko * Lukas Beckenbach * Thomas Göhrt *
Stefan Streif *

* *Technische Universität Chemnitz, Automatic Control and System Dynamics Laboratory, Germany (e-mail: p.osinenko@gmail.com, {patrick.schmidt, stefan.streif}@etit.tu-chemnitz.de)*

Abstract: Reinforcement learning (RL) in the context of control systems offers wide possibilities of controller adaptation. Given an infinite-horizon cost function, the so-called critic of RL approximates it with a neural net and sends this information to the controller (called "actor"). However, the issue of closed-loop stability under an RL-method is still not fully addressed. Since the critic delivers merely an approximation to the value function of the corresponding infinite-horizon problem, no guarantee can be given in general as to whether the actor's actions stabilize the system. Different approaches to this issue exist. The current work offers a particular one, which, starting with a (not necessarily smooth) control Lyapunov function (CLF), derives an online RL-scheme in such a way that practical semi-global stability property of the closed-loop can be established. The approach logically continues the work of the authors on parameterized controllers and Lyapunov-like constraints for RL, whereas the CLF now appears merely in one of the constraints of the control scheme. The analysis of the closed-loop behavior is done in a sample-and-hold (SH) manner thus offering a certain insight into the digital realization. The case study with a non-holonomic integrator shows the capabilities of the derived method to optimize the given cost function compared to a nominal stabilizing controller.

Keywords: Reinforcement learning control, Stability of nonlinear systems, Lyapunov methods

1. INTRODUCTION

Consider a general nonlinear dynamical system

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

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the input, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is the dynamics model. Further, consider the following infinite-horizon (IH) cost function:

$$J[\kappa](x_0) := \int_0^\infty \rho(x(t), \kappa(x(t)))\, \mathrm{d}t, \quad x(0) = x_0, \tag{2}$$

where $\rho : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ denotes the *reward* function and $\kappa : \mathbb{R}^n \to \mathbb{R}^m$ is a control policy. The function $J^*(x_0) := \min_\kappa J[\kappa](x_0), \forall x_0$ is called the *value function.* By the Bellman's optimality principle, it satisfies the Hamilton-Bellman-Jacobi equation:

$$\dot{J}^*(x) + \min_u \{\nabla J^* f(x, u) + \rho(x, u)\} = 0, \forall x. \tag{3}$$

Dynamic programming (DP) takes (3) as the basis, discretized (in a compact domain of) the state space, and computes and approximation to J^* in an iterative manner (Liu and Wei, 2014; Wei et al., 2016). The *curse of dimensionality* is what prevents application of DP in the dynamical context. One particular way to overcome this issue is to use parameterized function approximators $\hat{J}(x, \theta) = \langle \theta, \varphi(x) \rangle$ with a finite number of parameters, e.g. neural nets. Here, θ is the hidden layer weight vector and φ is the activation function of the net. Thus, the iterations are performed over the parameters θ. Roughly, the idea reads:

$$\begin{aligned} &\text{Step 1: } \theta_{\text{new}} := \min_\theta \{\rho + \Delta \hat{J}\} \quad \text{(Critic)}, \\ &\text{Step 2: } u_{\text{new}} := \min_u \{\rho + \hat{J}\} \quad \text{(Actor)}, \end{aligned} \tag{4}$$

where $\rho + \Delta \hat{J}$ represents the *Bellman error*, i. e., a metric, which describes the goodness of $\hat{J}$ as an approximation to J^* based on the HJB. There is a variety of RL methods, thus it is virtually impossible to comprehensively overview them (the reader may refer, e. g., to Bertsekas, 2017; Sutton and Barto, 2018; Recht, 2019). However, it is worthwhile to categorize some in terms of how they tackle the issue of closed-loop stability, since direct application of (4) does not necessarily give any guarantees.

There are methods that: (a) are heavily based on DP principles (Heydari, 2014; Wei et al., 2016), (b) concentrate solely on neural net weight learning (Sokolov et al., 2015; Zhang et al., 2011; Mu et al., 2017), (c) start with sufficiently good initial data (Jiang and Jiang, 2015; Gao and Jiang, 2017), (d) restrict to linear systems (Bian and Jiang, 2016; Gao and Jiang, 2016). The first category entails iterations over (a subset) of the state space, (b) require long off-line learning phases and do not take into account closed-loop stability, (c) puts the burden of fine initialization onto the user. In general, there is oftentimes a dilemma: RL pursues optimality of DP, but often lacks stability guarantees, whereas some nominal stabilizing controller is not concerned about optimality in the sense of minimizing (2). The relations between optimal and stabilizing controllers were well described in (Primbs et al., 1999). It seems a certain trade-off is required to tackle optimality and closed-loop stability simultaneously.

The **contribution** of the current work is to offer an RL-method which does address closed-loop stability. It is based on an initial stabilizability information, specifically, in the form of a (not necessarily smooth) CLF. The justification of such an assumption is as follows. Every existing RL approach requires at least stabilizability of the system. Stabilizability implies in turn existence of a CLF by a converse result. It is suggested to constrain the RL-method accordingly. Similar philosophy was pursued in the previous work of the authors (Beckenbach et al., 2018; Göhrt et al., 2019). However, the current work greatly generalizes the previous derivations. First, the assumed CLF needs not to be smooth, as it is in the general case (Clarke et al., 1997). Secondly, state convergence shown in this work is provided in the sense of practical stabilizability instead of just ultimate boundedness used some literature (see, e. g., Vamvoudakis and Lewis, 2010). The new algorithm is thus suggested in a sample-and-hold setting (SH) which gives insight into the digital realization. In particular, the actor and critic are now merged, and the "actor-critic" optimization is performed at discrete time samples. Roughly, the method reads:

$$\begin{aligned} (u_{\text{new}}, w_{\text{new}}) &:= \min_{(u,w)} \mathbb{J}(x,u,w) && \text{(Actor-Critic)}, \\ \text{s.t. } &\Delta_{\text{inter-sample}} \hat{J} < 0 && \text{(Constraints)}, \\ &\Delta_{\text{sample-to-sample}} \hat{J} < 0 \end{aligned} \tag{5}$$

where $\mathbb{J}$ is a cost function related to the Bellman error. Notice here the requirement of inter-sample and sample-to-sample decay of the critic $\hat{J}$. The actual algorithm, which is presented in Section 3, does not pose the constraints so literally – there are certain relaxation terms. The case study with a non-holonomic integrator demonstrates the worthiness of (5) in Section 4.

Notation. A closed ball of radius $R > 0$ centered at the origin is denoted $\mathcal{B}_R$. A continuous function $\alpha : [0, a) \to [0, \infty)$ is said to belong to class $\mathcal{K}$, if it is strictly increasing and $\alpha(0) = 0$. It is said to belong to $\mathcal{K}_\infty$ if $a = \infty$ and $\alpha(r) \to \infty$ as $r \to \infty$.

2. PRELIMINARIES

As mentioned above, the suggested RL-method will be considered in a SH setting. Such a setting means applying constant controls throughout sampling periods of some time $\delta > 0$, in which the system is governed by

$$\begin{aligned} &\dot{x} = f(x, u_k), \quad x(0), \\ &t \in [k\delta, (k+1)\delta), \ u_k = \kappa(x(k\delta)), \ k \in \mathbb{N}_0, \end{aligned} \tag{6}$$

where $u_k \in \mathbb{U} \subset \mathbb{R}^m$ are input constraints. It is assumed that the dynamics model f is locally Lipschitz in x for any $u \in \mathbb{U}$. In the following denote $x(k\delta) =: x_k$. For any $k \in \mathbb{N}_0$ and $\delta > 0$, the state $x^{u_k}(t)$ at $t \geq k\delta$ under u_k is defined as

$$x^{u_k}(t) := x_k + \int_{k\delta}^{t} f(x(\tau), u_k) \, \mathrm{d}\tau. \tag{7}$$

For $t = (k+1)\delta$, denote $x^{u_k}_{k+1} := x^{u_k}((k+1)\delta)$. The corresponding trajectory of (6) under the SH-mode input will also be called *SH-trajectory*. Consider the following standard

Definition 1. A control policy $\kappa(\cdot)$ is said to practically semi-globally stabilize (1) if, given $0 < r < R < \infty$, there exists a $\bar{\delta} > 0$ s. t. any SH-trajectory $x(t)$ with a sampling time $\delta \leq \bar{\delta}$, starting in $\mathcal{B}_R$ is bounded, enters $\mathcal{B}_r$ after a time T, which depends uniformly on R, r, and remains there for all $t \geq T$.

In the following, for brevity, the wording "semi-globally" is omitted. In the light of Definition 1, the balls $\mathcal{B}_R$ and $\mathcal{B}_r$ are denoted the *starting* and *target ball*, respectively. Recall further the following

Definition 2. For a locally Lipschitz function $V : \mathbb{R}^n \to \mathbb{R}$ and a $v \in \mathbb{R}^n$, the *generalized lower directional derivative* (GLDD) of V in the direction of v at x is defined as (Sontag and Sussmann, 1995)

$$\mathcal{D}_v V(x) := \liminf_{\tau \to 0_+} \frac{1}{\tau} \left(V(x + \tau v) - V(x) \right). \tag{8}$$

The following is a stabilizability assumption:

Assumption 1. There exists a locally Lipschitz continuous, positive-definite function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, a continuous positive definite function $w : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and $\alpha_{1,2} \in \mathcal{K}_\infty$ s. t. for any compact $\mathbb{X} \subset \mathbb{R}^n$, there exists a compact set $\mathbb{U}_\mathbb{X} \subseteq \mathbb{U}$ and it holds that, for any $x \in \mathbb{X}$,

i) V has a decay rate satisfying

$$\inf_{u \in \mathbb{U}_\mathbb{X}} \mathcal{D}_{f(x,u)} V(x) \leq -w(x), \tag{9}$$

ii) $\alpha_1(\|x\|) \leq V(x) \leq \alpha_2(\|x\|)$,

The pair (V, w) is also referred to as a *CLF-pair.*

Remark 1. Under the existence of a CLF as per Assumption 1, practical stabilization in the sense of SH can be realized as follows: given a CLF-pair (V, w) and balls with radii $0 < r < R < \infty$, there exists a $\bar{\delta} > 0$ s. t. for any $0 < \delta \leq \bar{\delta}$, there is a (possibly discontinuous) map $\mu : \mathbb{R}^n \to \mathbb{U}$ s. t. the SH-trajectory of (6) with the sampling period δ satisfies:

$$V\left(x^{\mu(x_k)}_{k+1}\right) - V(x_k) \leq -\frac{\delta}{2} w(x_k). \tag{10}$$

There are various methods of deriving a SH realization of μ (refer, e. g., to Clarke et al. (1997); Braun et al. (2017)). The meaning of the last displayed inter-sample decay condition is that, using (9), one may calculate such control actions $\mu(x_k)$ at the sampling nodes, that at least half the decay is retained (the relaxation comes from the inter-sample behavior).

Now, address the actor-critic setup of the paper. First, the critic

$$\hat{J} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}_{\geq 0} : (x, \theta) \mapsto \langle \theta, \varphi(x) \rangle, \tag{11}$$

can be regarded as a neural net, consisting of the hidden layer weights $\theta \in \mathbb{R}^p$ and a locally Lipschitz continuous activation function $\varphi : \mathbb{R}^n \to \mathbb{R}^p$. On one hand, given $\hat{J}$, it holds that on any compact $\mathbb{X} \subset \mathbb{R}^n$, $\hat{J}(x, \theta) \leq \|\theta\| L_\varphi \|x\|$, where L_φ is the corresponding local Lipschitz constant on $\mathbb{X}$. On the other hand, let the activation function satisfy the following condition: there exists $\underline{l} \in \mathcal{K}$ s. t. for any $x \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^p$, it holds that $\langle \theta, \varphi(x) \rangle \geq \underline{l}(\|x\|) \cdot \|\theta\|$.

Let the following assumption on the activation function of (11) and the CLF V as per Assumption 1 hold:

Assumption 2. There exists $\theta^{\#} \in \mathbb{R}^p$ s. t. $V(x) = \hat{J}(x, \theta^{\#})$, for all $x \in \mathbb{R}^n$.

Assumption 2 states that the structure of φ, which is a designer's choice, be "rich" enough to structurally capture V, which is known. It will be used in this form in the algorithm analysis of Section 3. Analogous structural assumptions can be found in, e. g., (Richards et al., 2018) to match desired properties of the parametric approximant. However, in principle, it may be relaxed to approximate structure matching without essential changes to the forthcoming analyses, and so is omitted for simplicity and brevity.

The actor-critic routine is suggested as follows. Given $0 < r < R$, consider the following optimization problem $\mathcal{AC}(x_k; R, r)$ at the state $x_k, k \in \mathbb{N}_0$, assuming $x(0) \in \mathcal{B}_R$:

$$\min_{(u,\theta)\in\mathbb{U}\times\Theta} \quad \mathbb{J}(x_k, u, \theta) \tag{A-C-Obj}$$

$$\text{s. t.} \quad \hat{J}(x_k, \theta) \leq \hat{J}(x_k, \theta_{k-1}) + \varepsilon_1 \tag{C1}$$

$$V(\hat{x}_{k+1}) \leq \hat{J}(\hat{x}^u_{k+1}, \theta) + \varepsilon_2 \tag{C2}$$

$$\hat{J}(\hat{x}^u_{k+1}, \theta) - \hat{J}(x_k, \theta) \leq -\frac{\delta}{2} w(x_k) + \varepsilon_3 \tag{C3}$$

$$q_1(\|x_k\|) \leq \hat{J}(x_k, \theta) \leq q_2(\|x_k\|). \tag{C4}$$

Here, the cost function is $\mathbb{J} : \mathbb{R}^n \times \mathbb{U} \times \mathbb{R}^p \to \mathbb{R}$, $\varepsilon_{1,2,3} \geq 0$, $q_{1,2} \in \mathcal{K}_\infty, \Theta \subset \mathbb{R}^p$ will be described down below, and $\hat{x}^u_{k+1}$ is the state prediction at $t = k\delta$ which is done via

$$\hat{x}^u_{k+1} = x_k + \delta f(x_k, u), \tag{13}$$

although other prediction schemes are possible. It is due to this state "prediction" that $\mathcal{AC}(x_k; R, r)$ incorporates relaxation terms $\varepsilon_{1,2,3}$. In (A-C-Obj), the objective function $\mathbb{J}$ is set so as to minimize the (squared) Bellman error:

$$\mathbb{J} = \Big(\rho(x_k, u) + \hat{J}(\hat{x}^u_{k+1}, \theta_{k-1}) - \hat{J}(x_k, \theta)\Big)^2. \tag{14}$$

Once the system's state x_k is in $\mathcal{B}_{r^*} \subset \mathcal{B}_r$ for some $r^* \leq r$, which is referred to as the *core ball*, the setting of (u_k, θ_k) is arbitrary. This is dictated by the fact that the optimization problem may become infeasible in a small vicinity of the origin due to the SH behavior. Nor is one interested in what happens there as far as SH-setting is concerned. Various variables in $\mathcal{AC}(x_k; R, r)$ as well as the dependence of the core ball size $r^* = r^*(R, r)$ on the starting and target balls is now described. First, regarding the positive-definiteness property of $\hat{J}$ as in (C4), construct $q_{1,2}$ by specifying bounds on the weight norm as follows. Let $\underline{\theta}, \overline{\theta}$ be s. t. $\underline{\theta} \leq \|\theta^{\#}\| \leq \overline{\theta}$ (see Assumption 2) and define

$$\Theta := \{\theta \in \mathbb{R}^p : \underline{\theta} \leq \|\theta\| \leq \overline{\theta}\}, \; q_1(\|x\|) := \underline{l}(\|x\|) \cdot \underline{\theta}. \tag{15}$$

Next, set

$$\bar{J} := \sup_{x\in\mathcal{B}_R, \theta\in\Theta} \hat{J}(x, \theta).$$

Fix an $\eta_R > 0$ and specify $R^* > R$ s. t. $q_1(R^*) \geq \bar{J} + \eta_R$. Provided with R^*, let $L_\varphi > 0$ be the local Lipschitz constant of φ on $\mathcal{B}_{R^*}$ and define

$$q_2(\|x\|) := \overline{\theta} L_\varphi \|x\|. \tag{16}$$

Furthermore, let $v^* = q_1(r)$ and $r^* = q_2^{-1}(\frac{v^*}{2})$ (the latter exists since q_2 is strictly increasing), which also implies $r^* \leq r$. At this point, note that for any $\theta \in \Theta$,

$$q_1(\|x\|) \leq \hat{J}(x, \theta) \leq v^* \;\Rightarrow\; \|x\| \leq r$$

and also

$$\frac{v^*}{2} \leq \hat{J}(x, \theta) \leq q_2(\|x\|) \;\Rightarrow\; \|x\| \geq r^*,$$

which relate the value of $\hat{J}$ to the facts that $x \in \mathcal{B}_r$ or $x \notin \mathcal{B}_{r^*(R,r)}$, respectively. It can be seen that, among other factors to be detailed later, the bounding functions $q_{1,2}$ contribute to the radius of the target ball.

Finally, call an actor-critic sequence $(u_k, \theta_k)_{k\in\mathbb{N}_0}$ *admissible* if, for any $k \in \mathbb{N}_0$, (C1)–(C4) are satisfied along the SH-trajectory of (6) as long as $x_k \notin \mathcal{B}_{r^*(R,r)}$. A single element of an actor-critic sequence will be called an *actor-critic tuple*. The following section is devoted to the analysis of the above optimization problem.

3. MAIN RESULTS

The following result presents necessary conditions under which $\mathcal{AC}(x_k; R, r), 0 < r < R$ yields a practically stabilizing control algorithm.

Theorem 2. Consider the control system (1) in the SH-mode (6) under the optimization $\mathcal{AC}(x_k; R, r)$. Let $q_{1,2}$ be according to (15) and (16), and $\bar{\delta}$ be defined as per Remark 1 for the radii $0 < r^* < R^*$. Assume that there exists an admissible actor-critic sequence $(u_k, \theta_k)_{k\in\mathbb{N}_0}$ for $\mathcal{AC}(x_k; R, r)$ along SH-trajectories of (6) with a sampling period $0 < \delta < \bar{\delta}$, and under some $\varepsilon_{1,2,3} \geq 0$. Then, there exist $0 < \bar{\delta}_0 \leq \bar{\delta}, \bar{\varepsilon}_{1,3} > 0$ with the following property: if the sampling period $\delta > 0$ satisfies $\delta \leq \bar{\delta}_0$ and $\varepsilon_1 \leq \bar{\varepsilon}_1, \varepsilon_3 \leq \bar{\varepsilon}_3$, then the control action sequence $u_k, k \in \mathbb{N}_0$ of the actor-critic sequence $(u_k, \theta_k)_{k\in\mathbb{N}_0}$ practically stabilizes the origin of (1).

Proof. First, let L_f be the Lipschitz constant of f on $\mathcal{B}_{R^*}$ and define

$$\bar{f} := \sup_{\substack{x\in\mathcal{B}_{R^*} \\ u\in\mathbb{U}}} f(x, u).$$

Now, let $(u_k, \theta_k)_{k\in\mathbb{N}_0}$ be an admissible actor-critic sequence. Consider (C1) at time $k + 1$:

$$\hat{J}(x^{u_k}_{k+1}, \theta_{k+1}) \leq \hat{J}(x^{u_k}_{k+1}, \theta_k) + \varepsilon_1,$$

where $x^{u_k}_{k+1}$ is the state after applying u_k at time $t = k\delta$. Using the fact that

$$\begin{aligned} \|x^u_{k+1} - \hat{x}^u_{k+1}\| &\leq \left\| \int_{k\delta}^{(k+1)\delta} f(x(\tau), u) \,\mathrm{d}\tau - f(x_k, u)\delta \right\| \\ &\leq \left\| \int_{k\delta}^{(k+1)\delta} f(x(\tau), u) - f(x_k, u) \,\mathrm{d}\tau \right\| \\ &\leq \int_{k\delta}^{(k+1)\delta} \|f(x(\tau), u) - f(x_k, u)\| \,\mathrm{d}\tau \\ &\leq \int_{k\delta}^{(k+1)\delta} L_f \bar{f}\delta \,\mathrm{d}\tau = L_f \bar{f} \delta^2 \end{aligned}$$

for any $k \in \mathbb{N}_0$, $u \in \mathbb{U}$ and $\delta > 0$, it holds that

$$\begin{aligned} \hat{J}(x^{u_k}_{k+1}, \theta_k) &\leq \hat{J}(\hat{x}^{u_k}_{k+1}, \theta_k) + L_\varphi \|\theta_k\| \left\|x^u_{k+1} - \hat{x}^u_{k+1}\right\| \\ &\leq \hat{J}(\hat{x}^{u_k}_{k+1}, \theta_k) + \overline{\theta} L_\varphi L_f \bar{f} \delta^2, \end{aligned} \tag{17}$$

for any $\theta_k \in \Theta$. Substituting this into (C1) at $k + 1$ gives

$$\hat{J}(x^{u_k}_{k+1}, \theta_{k+1}) \leq \hat{J}(\hat{x}^{u_k}_{k+1}, \theta_k) + \overline{\theta} L_\varphi L_f \bar{f} \delta^2 + \varepsilon_1$$

and further subtracting $\hat{J}(x_k, \theta_k)$ thereof yields, using (C3),

$$\begin{aligned}\hat{J}(x^{u_k}_{k+1}, \theta_{k+1}) - \hat{J}(x_k, \theta_k) \\ \leq -\frac{\delta}{2} w(x_k) + \bar{\theta} L_\varphi L_f \bar{f} \delta^2 + \varepsilon_1 + \varepsilon_3,\end{aligned} \tag{18}$$

for any $k \in \mathbb{N}_0$.

In the following it is checked whether any trajectory that starts inside the starting ball $x(0) \in \mathcal{B}_R$ is confined to $\mathcal{B}_{R^*}$ and converges to $\mathcal{B}_r$. Let $\hat{J}(x_0, \theta_0) \leq \bar{J}$, for any $x(0) \in \mathcal{B}_R$ and $\theta_0 \in \Theta$. Observe that

$$\hat{J}(x^{u_k}(t), \theta_k) \leq \hat{J}(\chi, \theta_k) + \bar{\theta} L_\varphi L_f \bar{f} \delta^2. \tag{19}$$

for either $\chi = x_k$ or $\chi = x^{u_k}_{k+1}$. Using (17) on the right-hand side above and subtracting $\hat{J}(x_k, w_k)$ on both sides, it holds that

$$\begin{aligned}\hat{J}(x^{u_k}(t), \theta_k) - \hat{J}(x_k, \theta_k) \\ \leq \hat{J}(\hat{x}^{u_k}_{k+1}, \theta_k) - \hat{J}(x_k, \theta_k) + 2\bar{\theta} L_\varphi L_f \bar{f} \delta^2,\end{aligned}$$

and furthermore, since the actor-critic tuple (u_k, θ_k) satisfies (C3),

$$\hat{J}(x^{u_k}(t), \theta_k) \leq \hat{J}(x_k, \theta_k) \underbrace{-\frac{\delta}{2} w(x_k) + 2\bar{\theta} L_\varphi L_f \bar{f} \delta^2 + \varepsilon_3}_{=:\Delta_k(\delta)}. \tag{20}$$

Then, under consideration of (19) at $k = 0$, δ need to satisfy $\bar{\theta} L_\varphi L_f \bar{f} \delta^2 \leq \eta_R$ as then

$$\hat{J}(x_0, \theta_0) + \bar{\theta} L_\varphi L_f \bar{f} \delta^2 \leq \bar{J} + \bar{\theta} L_\varphi L_f \bar{f} \delta^2 \leq q_1(R^*),$$

from which $\hat{J}(x^{u_0}(t), \theta_0) \leq q_1(R^*)$ and thus $\|x(t)\| \leq R^*$, $t \in [0, \delta)$, follows. However, $\hat{J}(x^{u_k}(t), \theta_k)$ can be upper bounded more strictly as in (20), for any $k \in \mathbb{N}_0$, from which the same conclusion follows if δ is s. t. $\Delta_0(\delta) \leq \eta_R$. Then, for all subsequent time steps $k \in \mathbb{N}_0$, boundedness of the SH-trajectory as $\|x(t)\| \leq R^*$, $t \in [k\delta, (k+1)\delta)$, follows if the value of $\hat{J}$ is non-increasing sample-wise, which is shown henceforth. Define now the minimal decay rate as

$$\bar{w} = \inf_{r^* \leq \|x\| \leq R^*} \frac{w(x)}{2},$$

by which $\Delta_k(\delta) = \Delta(\delta)$ can be made independent of the current state. In the following, it is shown that the right-hand side of (18) is strictly negative for any $k \in \mathbb{N}_0$ until $\hat{J}(x_k, \theta_k) \leq v^*$, i. e., that $\hat{J}$ decays to some limit sample-wise. Suppose that $\hat{J}(x_k, \theta_k) \geq \frac{v^*}{2}$: Under the minimal decay $\bar{w}$, (18) reads as

$$\begin{aligned}\hat{J}(x^{u_k}_{k+1}, \theta_{k+1}) - \hat{J}(x_k, \theta_k) \\ \leq -\delta \bar{w} + \bar{\theta} L_\varphi L_f \bar{f} \delta^2 + \varepsilon_1 + \varepsilon_3.\end{aligned} \tag{21}$$

Assume, that the right-hand side indeed is strictly negative, i. e., $\hat{J}(x^{u_k}_{k+1}, \theta_{k+1}) < \hat{J}(x_k, \theta_k)$. Then, at some time step $k \in \mathbb{N}$, the state enters the target ball $\mathcal{B}_r$ and furthermore reaches $\frac{v^*}{2} \leq \hat{J}(x_k, \theta_k) \leq \frac{3v^*}{4}$. Note that, by the Lipschitz property of the activation function,

$$\|x - y\| \leq \frac{1}{\bar{\theta} L_\varphi} \eta_r \;\Rightarrow\; |\hat{J}(x, \theta_k) - \hat{J}(y, \theta_k)| \leq \eta_r$$

for any $\eta_r > 0$ and $\theta_k \in \Theta$. Therefore, for $\|x^{u_k}(t) - x_k\| \leq \frac{1}{\bar{\theta} L_\varphi} \eta_r$, it holds that

$$\hat{J}(x^{u_k}(t), \theta_k) \leq \hat{J}(x_k, \theta_k) + \eta_r \leq \frac{3v^*}{4} + \eta_r.$$

This means, that $\hat{J}(x^{u_k}(t), \theta_k) \leq v^*$ and thus $\|x(t)\| \leq r$, $t \in [k\delta, (k+1)\delta)$, if δ satisfies $L_f \bar{f} \delta \leq \frac{1}{\bar{\theta} L_\varphi} \frac{v^*}{4}$. Therefore, choosing

$$\bar{\delta}_0 \leq \max_{0 < \delta \leq \bar{\delta}} \left\{ \delta \;\middle|\; \bar{\theta} L_\varphi L_f \bar{f} \delta \leq \frac{v^*}{4},\; \bar{\theta} L_\varphi L_f \bar{f} \delta^2 \leq \frac{\bar{w}}{10} \delta \right\} \tag{22}$$

and setting $\bar{\varepsilon}_1 := \frac{\bar{w}}{2} \delta$ and $\bar{\varepsilon}_3 := \frac{3\bar{w}}{10} \delta$, it follows that $x(t) \in \mathcal{B}_{R^*}$ due to (20), i. e., all SH-trajectories are bounded, and for $x_k \in \mathcal{B}_{R^*} \setminus \mathcal{B}_{r^*}$,

$$\hat{J}(x^{u_k}_{k+1}, \theta_{k+1}) - \hat{J}(x_k, \theta_k) \leq -\frac{\bar{w}}{10} \delta.$$

for any $0 < \delta \leq \bar{\delta}_0$. In (22), $\bar{\delta}$ represents a sampling time bound for V to have sample-wise decay on $\mathcal{B}_{R^*} \setminus \mathcal{B}_{r^*}$ (this will be made use of in Theorem 4).

Thus, it can be concluded that there exist a sampling time bound and relaxation terms of the actor-critic optimization problem s. t. (1) be practically stabilized in the SH-sense (6). The reaching time for the state to enter the target ball $\mathcal{B}_r$ can be determined in a uniform way from the decay rate and the value of $\hat{J}$ (see, e. g., Clarke et al., 1997; Osinenko et al., 2018). □

Remark 3. Observe that the sampling time bounds are lower for higher $\bar{\theta}$, which in turn is user-defined. More specifically, in (15), $\underline{\theta}$, $\bar{\theta}$ are design factors that influence the relation between the sampling time and the overshoot as well as the core ball, chosen s. t. $\theta^{\#} \in \Theta$.

While the previous result states that the system can be practically stabilized if the sampling time as well as the relaxation terms are chosen sufficiently small, it needs to be shown that indeed for all times $k \in \mathbb{N}_0$, there exists an admissible actor-critic sequence $(u_k, \theta_k)_{k \in \mathbb{N}_0}$ satisfying (C2)-(C4). For that matter, recall that Assumption 1 ensures the existence of a CLF while Assumption 2 establishes a structural richness $\hat{J}$ to capture V. Therefore, a nominal stabilizing control policy $\mu(\cdot)$ associated with the CLF V, along with $\theta^{\#}$, can guarantee existence of admissible actor-critic sequences for $\mathcal{AC}(x_k; R, r)$, as summarized in the following

Theorem 4. Let Assumption 1-2 hold. Let $\varepsilon_1 \leq \bar{\varepsilon}_1$, $\varepsilon_3 \leq \bar{\varepsilon}_3$ in $\mathcal{AC}(x_k; R, r)$ and $0 < \delta \leq \bar{\delta}_0$, $\theta \in \Theta$ be bounded as per Theorem 2. Given $0 < r < R$, with the corresponding $0 < r^* < R^*$, $\bar{w} > 0$, and $x(0) \in \mathcal{B}_R$, there exists $\underline{\varepsilon}_1 > 0$, $\bar{\varepsilon}_2 > 0$, $\underline{\varepsilon}_3 > 0$, $0 < \bar{\delta}_1 \leq \bar{\delta}_0$ s. t. the following holds: if $\underline{\varepsilon}_1 \leq \varepsilon_1 \leq \bar{\varepsilon}_1$, $0 \leq \varepsilon_2 \leq \bar{\varepsilon}_2$, $\underline{\varepsilon}_3 \leq \varepsilon_3 \leq \bar{\varepsilon}_3$ and $0 < \delta \leq \bar{\delta}_1$, then for all times $k \in \mathbb{N}_0$ where $x_k \notin \mathcal{B}_{r^*}$, there exists an admissible actor-critic tuple giving rise to an admissible actor-critic sequence $(u_k, \theta_k)_{k \in \mathbb{N}_0}$.

Proof. Consider a current state $x_k \in \mathcal{B}_{R^*} \setminus \mathcal{B}_{r^*}$ at some time step $k \in \mathbb{N}_0$. By Assumption 2, it holds that

i) $\hat{J}(x^{\mu(x_k)}_{k+1}, \theta^{\#}) - \hat{J}(x_k, \theta^{\#}) \leq -\delta \frac{w(x_k)}{2}$,
ii) $\alpha_1(\|x_k\|) \leq \hat{J}(x_k, \theta^{\#}) \leq \alpha_1(\|x_k\|)$,

where i) is attained for $0 < \delta \leq \bar{\delta}_0 \leq \bar{\delta}$ (which was satisfied by Theorem 2). Given these properties, it needs to be shown that (C1)-(C4) are well posed to mean that these constraints are feasible for all times where $x_k \notin \mathcal{B}_{r^*}$. From i), it follows that

$$\hat{J}(\hat{x}_{k+1}^{\mu(x_k)}, \theta^{\#}) - \hat{J}(x_k, \theta^{\#}) \leq -\delta \frac{w(x_k)}{2} + \|\theta^{\#}\| L_\varphi L_f \bar{f} \delta^2$$
$$\leq -\delta \bar{w} + \frac{\bar{w}}{10}\delta.$$

Hence, ε_3 is lower bounded to be at least $\underline{\varepsilon}_3 := \frac{\bar{w}}{10}\delta$, which results in

$$\underline{\varepsilon}_3 := \frac{\bar{w}}{10}\delta \leq \varepsilon_3 \leq \frac{3\bar{w}}{10}\delta =: \bar{\varepsilon}_3.$$

Constraint (C2) is satisfied for any $\varepsilon_2 \geq 0$ due to Assumption 2. Next, note that $\theta_k = \theta^{\#}$ satisfies (C1) only if the sum of the value of $\hat{J}$ under θ_{k-1} and of ε_1 is not less than the value of V. Observe, however, that $\hat{J}$ with θ_{k-1} is lower bounded due to (C2) as in

$$V(\hat{x}) \leq \hat{J}(\hat{x}_k^{u_{k-1}}, \theta_{k-1}) + \varepsilon_2,$$

from which it follows that

$$V(\hat{x}_k^{u_{k-1}}) \leq \hat{J}(x_k^{u_{k-1}}, \theta_{k-1}) + \varepsilon_2 + \bar{\theta} L_\varphi L_f \bar{f} \delta^2$$

and thus $V(x_k)$ satisfies

$$V(x_k^{u_{k-1}}) \leq \hat{J}(x_k^{u_{k-1}}, \theta_{k-1}) + \varepsilon_2 + \underbrace{\bar{\theta} L_\varphi L_f \bar{f} \delta^2}_{\leq \frac{\bar{w}}{10}\delta} + L_V L_f \bar{f} \delta^2. \quad (23)$$

Hence it needs to be shown that $\hat{J}(x_k, \theta^{\#}) = V(x_k)$ is feasible in (C1) for $\varepsilon_1 \leq \bar{\varepsilon}_1$, that arose from the stability requirements, given the fact that (23) holds. Conversely, since from (C1)

$$\hat{J}(x_k^{u_{k-1}}, \theta^{\#}) \leq \hat{J}(x_k^{u_{k-1}}, \theta_{k-1}) + \varepsilon_1,$$

ε_1 is lower bounded by the second line of (23) as

$$0 \leq \varepsilon_2 + \frac{\bar{w}}{10}\delta + L_V L_f \bar{f} \delta^2 \leq \varepsilon_1 \leq \bar{\varepsilon}_1.$$

Setting

$$0 \leq \varepsilon_2 \leq \frac{\bar{w}}{10}\delta =: \bar{\varepsilon}_2$$

and constraining the sample time s. t. $0 < \delta \leq \bar{\delta}_1$, in which

$$\bar{\delta}_1 \leq \max_{0 \leq \delta \leq \bar{\delta}_0} \left\{ \delta \;\middle|\; L_V L_f \bar{f} \delta^2 \leq \frac{\bar{w}}{10}\delta \right\},$$

results in

$$\underline{\varepsilon}_1 := \frac{3\bar{w}}{10}\delta \leq \varepsilon_1 \leq \frac{\bar{w}}{2}\delta =: \bar{\varepsilon}_1,$$

under which $\theta_k = \theta^{\#}$ is admissible for (C1), for any $0 < \delta \leq \bar{\delta}_1$. Finally, regarding (C4), $q_{1,2}$ were chosen according to the specified weight norm bounds $\underline{\theta}, \bar{\theta}$, which allowed $\theta^{\#} \in \Theta$. Hence, feasibility of $\mathcal{AC}(x_k; R, r)$ is shown. □

Remark 5. Note that Assumption 1-2 are only necessary in Theorem 4 in order to serve feasibility in Theorem 2.

Remark 6. Due to Assumption 2, the Lipschitz constants L_V and L_φ are related. They may be comprised to $\bar{L} = \max\{L_V, L_\varphi\}$ in Theorem 4. However, this may lead to a tighter restriction on the sampling time bounds.

The above results can be summarized in the following

Theorem 7. Consider the control system (1) in the SH-mode (6) under the actor-critic optimization $\mathcal{AC}(x_k; R, r)$. Let Assumption 1–2 hold and let $\bar{\delta}$ be as per Remark 1 for radii $0 < r^* < R^*$. Then, there exist bounds $\underline{\varepsilon}_{1,2,3}, \bar{\varepsilon}_{1,2,3} \geq 0$ and $0 < \bar{\delta}_1 \leq \bar{\delta}$ s. t. for any $x(0) \in \mathcal{B}_R$, there exists an admissible actor-critic sequence, the action sequence of which practically stabilizes (1) as per Definition 1, if $\underline{\varepsilon}_i \leq \varepsilon_i \leq \bar{\varepsilon}_i$, $i = 1, 2, 3$, and $0 < \delta \leq \bar{\delta}_1$.

4. CASE STUDY AND DISCUSSION

In the following, the non-holonomic integrator

$$(\dot{x}_1, \dot{x}_2, \dot{x}_3)^\top = (u_1, u_2, x_1 u_2 - x_2 u_1)^\top$$

is to be practically stabilized under the presented approach. It was shown in, e. g., (Clarke, 2011), that the function

$$V(x) = x_1^2 + x_2^2 + 2x_3^2 - 2|x_3|\sqrt{x_1^2 + x_2^2}$$

is a global CLF for the system under the input constraint $u \in [-1, 1]^2$. In the case study, a nominal practically stabilizing control policy is computed via an Inf-convolution technique discussed in (Clarke et al., 1997; Osinenko et al., 2018). In the suggested RL-method, the activation function is set to

$$\varphi = \left(x_1^2, x_2^2, 2x_3^2, -2|x_3|\sqrt{x_1^2 + x_2^2} \right)^\top,$$

so that $\theta^{\#} = (1, 1, 1, 1)^\top$. In (2) as well as in (14), $\rho(x, u) = 0.1 x^\top x + 2 u^\top u$. First, the constraints (C1)–(C3) are relaxed with $\varepsilon_{1,2,3} = 5 \cdot 10^{-8}$. The sampling time is set to $\delta = 0.01$ and the radius of the target ball is set to $r = 0.1$. The trajectory of (1) under the suggested RL-method in the SH-mode, starting at $x(0) = (-2, -1.5, 0.4)^\top$ can be seen in Fig. 1.

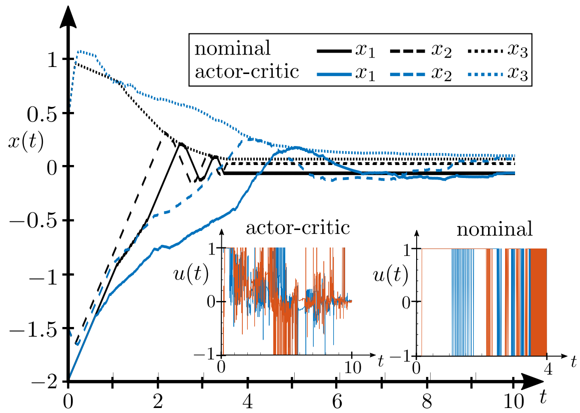

Fig. 1. State trajectory under the RL-method (blue) and the nominal controller (black) with the corresponding controls (u_1 blue, u_2 red). While a similar, oscillatory-like pattern can be detected in both state trajectories, that under the RL-method converges slower to the target. Yet, the action effort u reaches its constraint boundaries less often compared to the nominal policy.

From Fig. 1 it can be deduced that stabilization occurs under less action effort whilst allowing the state to converge slower to the target. Certain approaches, e. g., steepest descent, were observed to lead to a bang-bang control chattering between the borders of the set $\mathbb{U}$ (refer to, e. g., Braun et al., 2017; Osinenko et al., 2018), whereas such a behavior could be somewhat alleviated by using the suggested method. Consider the simulated, quasi-IH cost

$$J_{\text{sim}}[(u_k)_{k \in \mathbb{N}_0}](x_0) = \sum_{k=0}^{T-1} \int_{k\delta}^{(k+1)\delta} \rho(x(t), u_k)\, \mathrm{d}t$$

under a control sequence $(u_k)_{k \in \mathbb{N}_0}$, where T is the reaching time of the ball $\mathcal{B}_r$ from the starting ball $\mathcal{B}_R$ with $R = 1.75$, i. e., the cost of driving the state from $x(0) \in \mathcal{B}_R$ to $\mathcal{B}_r$. Using fixed $x_3(0) = 0.4$, the cost difference percentage

$$J_{\mathrm{sim},\%}(x_0) = \frac{J_{\mathrm{sim}}[(u_k)_{k\in N_0}^{\text{actor-critic}}](x_0)}{J_{\mathrm{sim}}[(u_k)_{k\in N_0}^{\text{nominal}}](x_0)} \cdot 100\%$$

on a $(x_1(0), x_2(0)) \in [-1.2,\ 1.2]^2$ grid is depicted in Fig. 2.

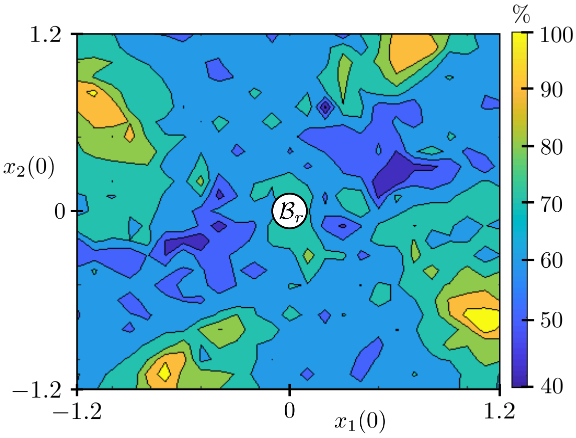

Fig. 2. Contour of $J_{\mathrm{sim},\%}(x_0)$. The contour captures the compared cost of driving the state from an initial state in the domain $[-1.2, 1.2]^2 \times 0.4$ to a target ball. In most of the grid, the actor-critic control policy could reduce the cost by $20 - 40\%$.

It can be seen that the quasi-IH cost under the RL-method could be improved significantly over a nominal controller.

5. CONCLUSION

This work was concerned with closed-loop stability issues of RL-methods for dynamical systems. It suggested to use an initial stabilizability information, specifically, a (non-smooth) CLF, and to introduce it into the constraints of the control scheme. Practical semi-global stabilizability of the closed-loop, resulting from application of the new method, is analyzed in sample-and-hold manner, which in turn gives insight into the digital realization. The case study with a non-holonomic integrator showed the merit of the new ideas.

REFERENCES

Beckenbach, L., Osinenko, P., and Streif, S. (2018). Addressing infinite-horizon optimality in MPC via Q-learning. *IFAC-PapersOnLine*, 51(20), 60–65.

Bertsekas, D.P. (2017). Value and policy iteration in optimal control and adaptive dynamic programming. *IEEE Trans. Neural Netw. Learn. Sys.*, 28(3), 500 – 509.

Bian, T. and Jiang, Z.P. (2016). Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. *Automatica*, 71, 348–360.

Braun, P., Grüne, L., and Kellett, C.M. (2017). Feedback design using nonsmooth control Lyapunov functions: A numerical case study for the nonholonomic integrator. In *Proc. of the 56th IEEE Conf. on Decision and Control*.

Clarke, F. (2011). Lyapunov functions and discontinuous stabilizing feedback. *Annual Rev. Control*, 35(1), 13–33.

Clarke, F., Ledyaev, Y., Sontag, E., and Subbotin, A. (1997). Asymptotic controllability implies feedback stabilization. *IEEE Trans. Automat. Control*, 42(10), 1394–1407.

Gao, W. and Jiang, Z.P. (2016). Adaptive Dynamic Programming and adaptive optimal output regulation of linear systems. *IEEE Trans. Automat. Control*, 61(12), 4164–4169.

Gao, W. and Jiang, Z.P. (2017). Nonlinear and adaptive suboptimal control of connected vehicles: A global Adaptive Dynamic Programming approach. *J. Intelligent & Robotic Syst.*, 85(3-4), 597–611.

Göhrt, T., Osinenko, P., and Streif, S. (2019). Adaptive dynamic programming using lyapunov function constraints. *IEEE Control Syst. Lett.*, 3(4).

Heydari, A. (2014). Revisiting approximate dynamic programming and its convergence. *IEEE Trans. on Cyb.*, 44(12), 2733–2743.

Jiang, Y. and Jiang, Z.P. (2015). Global adaptive dynamic programming for continuous-time nonlinear systems. *IEEE Trans. Automat. Control*, 60(11), 2917–2929.

Liu, D. and Wei, Q. (2014). Policy Iteration adaptive dynamic programming algorithm for discrete-timeime nonlinear systems. *IEEE Trans. Neural Netw. Learn. Syst.*, 25(3), 621–634.

Mu, C., Wang, D., and He, H. (2017). Novel iterative neural dynamic programming for data-based approximate optimal control design. *Automatica*, 81, 240–252.

Osinenko, P., Beckenbach, L., and Streif, S. (2018). Practical Sample-and-Hold Stabilization of Nonlinear Systems Under Approximate Optimizers. *IEEE Control Syst. Lett.*, 2(4), 569–574.

Primbs, J.A., Nevistić, V., and Doyle, J.C. (1999). Nonlinear optimal control: A control Lyapunov function and receding horizon perspective. *Asian J. Control*, 1(1), 14–24.

Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. *Annual Rev. Control, Robotics and Autonom. Syst.*, 2, 253–279.

Richards, S.M., Berkenkamp, F., and Krause, A. (2018). The Lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems. Available at arXiv:1808.00924v2 [cs.SY].

Sokolov, Y., Kozma, R., Werbos, L.D., and Werbos, P.J. (2015). Complete stability analysis of a heuristic approximate dynamic programming control design. *Automatica*, 59, 9–18.

Sontag, E.D. and Sussmann, H.J. (1995). Nonsmooth control-lyapunov functions. In *Proc. of the 34th IEEE Conf. on Decision and Control*.

Sutton, R.S. and Barto, A.G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, 2nd ed. edition.

Vamvoudakis, K.G. and Lewis, F.L. (2010). Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46(5), 878–888.

Wei, Q., Liu, D., and Lin, Q. (2016). Discrete-time local Value Iteration Adaptive Dynamic Programming: Convergence analysis. *IEEE Trans. Syst. Man Cyb.*, 48(6), 875–891.

Zhang, H., Cui, L., and Luo, Y. (2011). Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using Adaptive Dynamic Programming method. *IEEE Trans. Neural Netw.*, 22(12), 2226–2236.