A reinforcement learning method with closed-loop stability guarantee for systems with unknown parameters

Thomas Göhrt [*] **Fritjof Griesing-Scheiwe** [*] **Pavel Osinenko** [*]
Stefan Streif [*]

[*] *Technische Universität Chemnitz, Automatic Control and System Dynamics Lab, 09107 Chemnitz, Germany (e-mails: fritjof.griesing-scheiwe, pavel.osinenko, thomas.goehrt, stefan.streif@etit.tu-chemnitz.de)*

Abstract: This work is concerned with the application of reinforcement learning (RL) techniques to adaptive dynamic programming (ADP) for systems with partly unknown models. In ADP, one seeks to approximate an optimal infinite horizon cost function, the value function. Such an approximation, i. e., critic, does not in general yield a stabilizing control policies, i. e., stabilizing actors. Guaranteeing stability of nonlinear systems under RL/ADP is still an open issue. In this work, it is suggested to use a stability constraint directly in the actor-critic structure. The system model considered in this work is assumed to be only partially known, specifically, it contains an unknown parameter vector. A suitable stabilizability assumption for such systems is an adaptive Lyapunov function, which is commonly assumed in adaptive control. The current approach formulates a stability constraint based on an adaptive Lyapunov function to ensure closed-loop stability. Convergence of the actor and critic parameters in a suitable sense is shown. A case study demonstrates how the suggested algorithm preserves closed-loop stability, while at the same time improving an infinite-horizon performance.

Keywords: Consensus and Reinforcement learning control, Nonlinear adaptive control

1. INTRODUCTION

In ADP, one is commonly concerned with an infinite-horizon (IH) optimal control problem for nonlinear systems of the form

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

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the input, $f : \mathbb{R}^n \to \mathbb{R}^n$ is the so called internal dynamics model and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ is called the input coupling function. The IH cost function is usually given as

$$J_\kappa(x_0) := \int\limits_0^\infty r(x(t), \kappa(x(t)))\mathrm{d}t, \quad x(0) = x_0, \tag{2}$$

where $r : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ 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*, and, by the Bellman's optimality principle (Bellman, 1957), satisfies the Hamilton-Bellman-Jacobi (HJB) equation (Lewis et al., 2012)

$$\min_u(\nabla_x J^* f(x,u) + r(x,u)) = 0, \forall x \in \mathbb{R}^n. \tag{3}$$

where $\nabla_x J$ is the usual nabla operator. Dynamic programming (DP) bases upon (3), discretizes (a compact domain of) the state space, and computes an approximation to J^* in an iterative manner over the said discretization nodes (Liu and Wei, 2014; Wei et al., 2016; Bertsekas and Tsitsiklis, 1995). The *curse of dimensionality*, which is related to the combinatorial explosion of the node quantity as the state dimension grows, prevents application of DP in an online manner.

In ADP, it is usually suggested to use function approximators $\hat{J}(\theta, x)$ for the unknown value function, e. g., represented by neural networks (Liu et al., 2017). However, it is not fully known how usage of the approximator $\hat{J}$ in the closed control loop affects stability (Balakrishnan et al., 2008; Sokolov et al., 2015). Furthermore, although it is often times claimed that ADP effectively deals with systems with unknown dynamics models, in fact, many approaches explicitly or tacitly use at least some model knowledge, e. g., that of the input-coupling function (Lewis and Vrabie, 2009; Liu and Wei, 2014).

The current work aims to address systems with partially known parameters. Namely, the following system class is considered here:

$$\dot{x} = f(x) + \Psi(x)\theta + g(x)u, \tag{4}$$

where $\theta \in \mathbb{R}^b$ is a constant, unknown parameter. Further, $f : \mathbb{R}^n \to \mathbb{R}^n, g : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^m$, and $\Psi : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^b$ are assumed known, smooth functions with $f(0) = 0$ and $\Psi(0) = 0$.

The major goal of the work is to develop an actor-critic structure for (4) that deals with the unknown parameter θ and establishes a closed-loop stability guarantee. The form (4) is commonly used in adaptive control (Krstić et al., 1995). Adaptive control methods usually suggest concrete control policies for stabilization, based on auxiliary machinery of tuning functions, parameter adaptation rules etc. One of the central assumptions is the knowledge of an adaptive control Lyapunov function (ACLF), which, roughly speaking, features a stabilizing control for (4) under perfect knowledge of θ. Adaptive control designs are then used to derive a realizable policy out of this ACLF by a suitable parameter adaptation $\hat{\theta}$ (Krstić et al., 1995). In the current work, it is suggested also to use an ACLF, but not to derive a concrete stabilizing policy, but rather to integrate it into an actor-critic structure in the form of a constraint and derive a stabilizing dynamic controller (see details in Section 3). This works extends the approach of Göhrt et al. (2019a,b) to systems with partially unknown dynamics of the form (4).

The rest of the paper is organized as follows. Section 2 introduces necessary definitions and notation. The actor-critic structure is as well as the update rules are introduced in Section 3. In Section 4, the proposed optimization problems for the actor and critic are analyzed with respect to convexity, feasibility and convergence to prescribed vicinities of the respective optima. Based on the feasibility for all time, stability under the approximate control policy is shown. The analysis section is followed by case study in Section 5.

2. PRELIMINARIES

This section introduces required notation and definitions. The partial derivative of a function is abbreviated as $\frac{\partial V}{\partial x} = V_x$. The norm $\|\cdot\|$ denotes the usual Euclidean norm of a vector or the spectral norm for matrices, respectively. The set $\mathbb{R}^+$ denotes the set of positive real numbers, $\mathbb{R}_0^+$ the set of positive real numbers including zero and $\mathbb{R}^-$ the negative real numbers including zero. Function arguments are omitted from here on to simplify presentation and are only provided at the first instance or when necessary.

The starting point of this work is the system (4) with partially unknown model. As already mentioned in the introduction, such a system description with a linear dependence on an unknown parameter is often used in adaptive control. Stability in this context can be studied, e. g., by using an ACLF (Krstić et al., 1995) defined as follows

Definition 1. (Adaptive control Lyapunov function).
A function $W : \mathbb{R}^n \times \mathbb{R}^b \to \mathbb{R}_0^+$ is called an adaptive control Lyapunov function for a system (4), if there exists a positive-definite matrix $\Gamma \in \mathbb{R}^{b\times b}$ such that for each $\theta \in \mathbb{R}^b$, $W(x,\theta)$ is a CLF for the modified system

$$\dot{x} = f(x) + \Psi(x)\left(\theta + \Gamma W_\theta(x,\theta)\right) + g(x)u, \tag{5}$$

in other words, W satisfies

$$\begin{aligned}\inf_{u\in\mathbb{R}} \big(W_x^\top(x,\theta)\left(f(x) + \Psi(x)\left(\theta + \Gamma W_\theta(x,\theta)\right)\right) + \\ g(x)u)\big\} < 0.\end{aligned} \tag{6}$$

ACLFs arise in various applications, e. g., in traction control (Nakakuki et al., 2008), where the structural physical description of the vehicle dynamics is known, but concrete wheel-ground parameters are not. In the current work, such an ACLF for the system (4) is exploited by the following

Assumption 2. There exists a continuously differentiable control policy $v(x,\theta)$ and a twice continuously differentiable radially unbounded ACLF $W : \mathbb{R}^n \times \mathbb{R}^b \to \mathbb{R}$ such that for all $x \in \mathbb{R}^n \setminus \{0\}$ and for all $\theta \in \mathbb{R}^b$, it holds that

$$\dot{W} = W_x^\top\left(f + \Psi(\theta + \Gamma W_\theta) + gv(x,\theta)\right) \le -\nu(x), \tag{7}$$

where $\nu : \mathbb{R}^n \to \mathbb{R}_0^+$ is a radially unbounded positive definite continuously differentiable function.

Notice that the control policy in this assumption is based on the unknown parameter θ.

The convergence analysis requires the notion of persistence of excitation, which is a common assumption in adaptive control (Krstić et al., 1995). This work uses the definition of PE defined in (Göhrt et al., 2019b):

Definition 3. (Persistence of excitation). A continuous matrix valued function A of time t is called uniformly T-persistently excited at level δ if, for all t the matrix

$$\int\limits_{t-T}^{t} A(\tau)\mathrm{d}\tau \ - \delta I \tag{8}$$

is positive-semidefinite. Here, I denotes the identity matrix.

Definition 4. (m-strongly convex). A real valued twice continuously differentiable function $J : D \to \mathbb{R}$ is called m-strongly convex, if there exists a $m \in \mathbb{R}^+$, such that $J_{xx}(x) - mI$ is positive-semidefinite for all $x \in D$, where J_{xx} denotes the Hessian of J.

The reward r of the IH problem (2) is assumed in the following form:

$$r(x,u) := q(x) + \rho(u), \tag{9}$$

where $q(0) = 0$ and $\rho(0) = 0$. It is assumed that $\rho(u) \in \mathbb{R}^+$ for all $u \neq 0$ and $q(x) \in \mathbb{R}^+$ for all $x \neq 0$. Furthermore, both functions are twice continuously differentiable and ρ be ζ-strongly convex.

The critic is suggested in the followig form

$$\hat{J}(x, w(t)) := w^\top(t)\varphi(x), \tag{10}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ is called activation function and assumed to be continuously differentiable and $w \in \mathbb{R}^d$ is a parameter vector. The critic is a parametrized function approximator for the IH cost given in (2).

The formulation of the actor-critic structure requires the notion of a

Definition 5. (Barrier-function). A twice continuously differentiable convex function $B : \mathbb{R}_0^- \to \mathbb{R} \cup \{\infty\}$ is called a *barrier function*, if $B(z) < \infty$ for all $z < 0$ and $\lim\limits_{z\to 0^-} B(z) \to \infty$.

The next section describes the suggested actor-critic structure for adaptively stabilizable system with partially unknown model.

3. ACTOR-CRITIC STRUCTURE

This section introduces the actor and critic optimization problems and the corresponding update rules. First, consider the actor optimization problem.

$$\begin{aligned} \min_{u} \quad & \Sigma(x,u,w) \\ \text{s.t.} \quad & \dot{x} = f(x) + \Psi(x)\theta + g(x)u, \\ & \dot{V}(x,\hat{\theta}) \leq -\nu(x), \end{aligned} \tag{11}$$

where

$$\Sigma(x,u,w) := r(x,u) + \underbrace{w^{\top}\dot{\varphi}}_{\hat{J}_x^{\top}\dot{x}}. \tag{12}$$

The cost function Σ is known as the Bellman error and is a common choice in ADP (Bertsekas and Tsitsiklis, 1995; Lewis and Vrabie, 2009). In particular, it recovers the HJB (3) under the approximated value function.

The function V in the inequality constraint of (11) is based on the ACLF theory (cf. (Krstić et al., 1995)) and is defined as

$$V(x,\hat{\theta}) := W(x,\hat{\theta}) + \frac{1}{2}\tilde{\theta}^{\top}\Gamma^{-1}\tilde{\theta}. \tag{13}$$

Here, $\hat{\theta}$ is an estimate of the unknown parameter θ and $\tilde{\theta} := \theta - \hat{\theta}$. The function W is the ACLF from Assumption 2.

The function $\dot{V}$ depends on the unknown parameter θ. This dependence can be removed as shown in (cf. (Krstić et al., 1995)[Theorem 4.3])

Lemma 6. Consider system (4). Let Assumption 2 hold. If

$$\dot{\hat{\theta}} = \Gamma\Psi^{\top}(x)W_x(x,\hat{\theta}), \tag{14}$$

then the time derivative of the Lyapunov function candidate V given in (13) can be made independent of θ.

Proof. The time derivative of V reads as

$$\dot{V} = W_x^{\top}\dot{x} + W_{\hat{\theta}}^{\top}\dot{\hat{\theta}} + \tilde{\theta}^{\top}\Gamma^{-1}\dot{\hat{\theta}}. \tag{15}$$

After applying the suggested adaptation rule, the derivative of the Lyapunov function candidate reads: (13):

$$\dot{V} = W_x^{\top}(f + \Psi\theta + gv) + W_{\hat{\theta}}^{\top}\Gamma\Psi^{\top}W_x + \tilde{\theta}^{\top}\Psi^{\top}W_x. \tag{16}$$

Using $\theta = \tilde{\theta} + \hat{\theta}$ and adding $W_x^{\top}(\Psi\Gamma W_{\hat{\theta}} - \Psi\Gamma W_{\hat{\theta}})$, yields

$$\begin{aligned} \dot{V} =& W_x^{\top}(f + \Psi(\hat{\theta} + \Gamma W_{\hat{\theta}}) + gv + \Psi\tilde{\theta} - \Psi\Gamma W_{\hat{\theta}}) \\ &+ W_{\hat{\theta}}^{\top}\Gamma\Psi^{\top}W_x - \tilde{\theta}^{\top}\Psi^{\top}W_x. \end{aligned} \tag{17}$$

After rearranging, one obtains

$$\dot{V} = W_x^{\top}(f + \Psi(\hat{\theta} + \Gamma W_{\hat{\theta}}) + gv), \tag{18}$$

which does not depend on θ. Due to Assumption 2, there exists a control policy $v(x,\theta)$ for each $\theta \in \mathbb{R}^b$ that guarantees (7). Since this includes $\hat{\theta}$ as well, there is a control $v(x,\hat{\theta})$ such that $\dot{V}(x,\hat{\theta},v(x,\hat{\theta})) < -\nu(x)$, which is independent of θ. ■

Using the result of Lemma 6, the actor optimization problem can be restated as

$$\begin{aligned} \min_{u} \quad & \Sigma(x,u,w) \\ \text{s.t.} \quad & \dot{x} = f + \Psi\theta + gu, \\ & \Lambda(x,\hat{\theta},u) \leq 0, \\ & \dot{\hat{\theta}} = \Gamma\Psi^{\top}W_x, \end{aligned} \tag{19}$$

where

$$\begin{aligned} \Lambda(x,\hat{\theta},u) :=& \dot{V}(x,\hat{\theta},u) + \nu(x) \\ =& W_x^{\top}(x,\hat{\theta})(f(x) + \Psi(x)(\hat{\theta} + \Gamma W_{\hat{\theta}}(x,\hat{\theta})) \\ &+ g(x)u) + \nu(x). \end{aligned} \tag{20}$$

This inequality constraint $\Lambda \leq 0$ is incorporated into the cost function via a barrier function.

$$\begin{aligned} \min_{u} \quad & \Pi(x,u,w,\hat{\theta},t) \\ \text{s.t.} \quad & \dot{x} = f + \theta\Psi + gu, \\ & \dot{\hat{\theta}} = \Gamma\Psi^{\top}W_x, \end{aligned} \tag{21}$$

where

$$\Pi(x,u,w,\hat{\theta},t) := \Sigma + \mu(t)B(\Lambda - \gamma), \tag{22}$$

with μ satisfying $\mu(t) > 0$ and $\dot{\mu}(t) < 0$ for all $t \geq 0$. One particular choice could be, e. g., $\dot{\mu} = -\mu$ with $\mu(0) = \mu_0 < \bar{\mu}$. The introduction of μ is motivated from discrete time interior point algorithms and is transfered to the continuous time case (Fazlyab et al., 2016). Furthermore, $\gamma > 0$ is a constant relaxation parameter, which prevents numerical issues with barrier functions, as convergence of the state would yield $\dot{V} \to 0$ leading to $B(z) \to \infty$.

Now the critic optimization problem is addressed.

$$\begin{aligned} \min_{w} \quad & \Upsilon(x_T, u_T, w) \\ \text{s.t.} \quad & \dot{x} = f + \theta\Psi + gu, \end{aligned} \tag{23}$$

where

$$\Upsilon(x_T, u_T, w) := \int\limits_{t-T}^{t} \Sigma^2(x(\tau), u(\tau), w(t))\mathrm{d}\tau. \tag{24}$$

The notation x_T and u_T are the state trajectory and input trajectory in the moving time interval $[t-T, t]$. Notice that w depends on t and not on τ in the integral, i. e., w is regarded as constant throughout the backward window. The right-hand side of (23) is the so called *interval reinforcement form* (Lewis and Vrabie, 2009).

In the following section, the update rules for the actor (21) and critic optimization problem (23) are introduced.

3.1 Update rules

The considered optimization problems are time-varying since the optimization occurs along the trajectory of the system. The asymptotic convergence to the, in general, time-varying optima require the inverse Hessian (Fazlyab et al., 2016). Depending on the considered problem, calculating the inverse Hessian might be computationally expensive. In order to avoid expensive calculations, the update rule for the actor optimization problem (21) is suggested to be simply gradient descent with time-varying gain, i. e.,

$$\dot{u} := -\alpha(t)\Pi_u(x,u,w,\hat{\theta},t), u(0) = v(x(0),\hat{\theta}(0)). \tag{25}$$

In case of the critic optimization problem, it is also suggested to use gradient descent, i. e.,

$$\dot{w} := -\beta(t)\Upsilon_w(x_T, u_T, w). \tag{26}$$

The optimal control action is defined as

$$\begin{aligned} u^*(t) := \arg\min_{u} \quad & \Pi(x,u,w,\hat{\theta},t) \\ \text{s.t.} \quad & \dot{x} = f + \Psi\theta + gu \end{aligned} \tag{27}$$

and the optimal critic parameters are defined as

$$\begin{aligned} w^*(t) := \arg\min_{w} \quad & \Upsilon(x_T, u_T, w) \\ \text{s.t.} \quad & \dot{x} = f + \Psi\theta + gu. \end{aligned} \tag{28}$$

The next section deals with analysis of the suggested actor-critic approach. In particular, the achieved convergence results are specified, since asymptotic convergence to the optima cannot be guaranteed as explained above.

4. ALGORITHM ANALYSIS

This section is concerned with the analysis of the suggested update rules (25) and (26).

4.1 Convexity

First, convexity of the critic optimization problem (23) with respect to the parameter vector w is analyzed. The convergence analysis requires strongly convex functions. For the critic cost function Υ, one need to make

Assumption 7. The matrix $\dot{\varphi}(x)\dot{\varphi}^\top(x)$ is T-persistently excited at level δ_1.

Based on this assumption, one can prove the following

Lemma 8. (Göhrt et al., 2019b). Under Assumption 7, the cost function Υ of the optimization problem (23) is δ_1-strongly convex.

In case of the actor cost function Π, convexity follows from the definition of the reward, in particular from ρ, which is shown in the next lemma.

Lemma 9. (Göhrt et al., 2019b). Consider the reward function (9) and the optimization problem (21). Let Assumption 2 hold. The cost function Σ defined in (22) is ζ-strongly convex.

The proofs of Lemma 8 and 9 can be found in Göhrt et al. (2019b).

4.2 Feasibility analysis

The critic optimization problem (23) is unconstrained, hence feasibility is not an issue. The equality constraint is already incorporated in the sense that w evolves through the update rule along the trajectory of the system.

The actor optimization problem (21) contains an inequality constraint relaxed via a barrier function. In this case, feasibility needs to be addressed. Feasibility of u in this context is equivalent to the existence of a solution for all $t \in \mathbb{R}_0^+$.

Lemma 10. Consider the system (4), the optimization problem (21), the update rule for the control action (25) and the update rule for the parameter $\hat{\theta}$ (14). Let Assumption 2 hold. For every initial condition $x(0) \in \mathbb{R}^n$ and $\hat{\theta}(0) \in \mathbb{R}^b$, the solution $u(t)$ to (25) satisfies $u(t) \in \mathcal{H}(x(t), \hat{\theta}(t))$ for all $t \geq 0$, if $u(0) \in \mathcal{H}(x(0), \hat{\theta}(0))$, where $\mathcal{H}(x, \hat{\theta}) := \{u \in \mathbb{R}^m : \Lambda(x, \hat{\theta}, u) \leq \gamma\}$.

Proof. The proof is analogous to that in Göhrt et al. (2019b) by considering an augmented state vector $z = (x, \hat{\theta})$. ■

4.3 Stability analysis

The feasibility result of the last section is now used to conclude stability of the system (4) under the control action u as a solution to (25).

Theorem 11. Consider the system (4) under an initial condition $x(0) = x_0$. Let Assumption 2 hold. For any continuously differentiable input $u : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ that satisfies

$$\begin{aligned}\dot{V}(x, \hat{\theta}, u) + \nu(x) - \gamma = W_x^\top(x, \hat{\theta})(f(x) + \Psi(x) \\ (\hat{\theta} + \Gamma W_{\hat{\theta}}(x, \hat{\theta})) + g(x)u) + \nu(x) - \gamma \leq 0, \\ \forall t \in \mathbb{R}_{\geq 0}, x(0) = x_0 \qquad (29)\end{aligned}$$

along the trajectory of (4), the state x converges to the set $\mathcal{G} := \{x \in \mathbb{R}^n : \nu(x) \leq \gamma\}$ and remains there.

Proof. The function $\nu(x)$ is radially unbounded and positive-definite, hence, proper. Therefore, the set $\mathcal{G}$ is compact. Define $E := \{x \in \mathbb{R}^n : \dot{V} = 0\}$. Since for all $x \in \mathbb{R}^n \setminus \mathcal{G}$, one has $\dot{V} < 0$, it follows that $E \subseteq \mathcal{G}$. By Lasalle's invariance principle (Khalil and Grizzle, 2002)[Theorem 4.4], $x(t)$ converges to the largest invariant subset of E. Since $E \subseteq \mathcal{G}$, $x(t)$ converges to $\mathcal{G}$. ■

4.4 Convergence analysis

This section analysis the convergence property of the proposed update rules. As already mentioned in Section 3.1, time-varying optimization problem have time-varying optima in general. Convergence to and tracking of the optima require the inverse Hessian in case of strongly convex optimization problems (Fazlyab et al., 2016). Using Hessian-free gradient descent as suggested therefore does not yield asymptotic convergence and tracking. But using specifically chosen gains for gradient descent, convergence to prescribed vicinities can be guaranteed as stated by the following

Theorem 12. Consider the system (4), the optimization problem (21) with the update rule (25) and the optimization problem (23) with the update rule (26). Let Assumptions 2 and 7 hold. For any $\varepsilon_1, \varepsilon_2 \in \mathbb{R}^+$, there exist $\alpha(t)$ and $\beta(t)$, such that $\lim_{t \to \infty} \|u(t) - u^*(t)\| \leq \varepsilon_1$ and $\lim_{t \to \infty} \|w(t) - w^*(t)\| \leq \varepsilon_2$ hold.

Proof. The proof is along the lines of Göhrt et al. (2019b). First, define $\tilde{u} := u - u^*$. Consider the following positive-definite function

$$V_0(u, w) := \underbrace{\frac{1}{2}\Pi_u^\top \Pi_u}_{=:V_1(\Pi_u)} + \underbrace{\frac{1}{2}\Upsilon_w^\top \Upsilon_w}_{=:V_2(\Upsilon_w)}. \qquad (30)$$

Its derivative is given as

$$\dot{V}_0 = \underbrace{\Pi_u^\top \dot{\Pi}_u}_{=:\dot{V}_1} + \underbrace{\Upsilon_w^\top \dot{\Upsilon}_w}_{=:\dot{V}_2}. \qquad (31)$$

The analysis is started with $\dot{V}_1$,

$$\dot{V}_1 = \Pi_u^\top (\Pi_{uu}\dot{u} + \Pi_{ux}\dot{x} + \Pi_{u\hat{\theta}}\dot{\hat{\theta}} + \Pi_{uw}\dot{w} + \Pi_{ut}). \qquad (32)$$

Together with (25), this yields

$$\dot{V}_1 = -\alpha \Pi_u^\top \Pi_{uu} \Pi_u + \Pi_u^\top (\Pi_{ux}\dot{x} + \Pi_{u\hat{\theta}}\dot{\hat{\theta}} + \Pi_{uw}\dot{w} + \Pi_{ut}). \qquad (33)$$

Since Π is ζ-strongly convex due to Lemma 9, it holds that

$$x^\top \Pi_{uu} x \geq \zeta x^\top x \qquad (34)$$

and further (Nesterov, 2013)

$$\Pi^* \geq \Pi + \Pi_u^\top (u^* - u) + \frac{\zeta}{2}\|\tilde{u}\|^2. \qquad (35)$$

Rewriting yields

$$\Pi_u^\top \tilde{u} \geq \underbrace{\Pi - \Pi^*}_{\geq 0} + \frac{\zeta}{2} \left\|\tilde{u}\right\|^2 \geq \frac{\zeta}{2} \left\|\tilde{u}\right\|^2 \geq 0. \tag{36}$$

It follows that

$$\frac{\zeta^2}{4} \left\|\tilde{u}\right\|^4 \leq (\Pi_u^\top \tilde{u})^2 \leq \left\|\Pi_u\right\|^2 \left\|\tilde{u}\right\|^2, \tag{37}$$

and, therefore,

$$\frac{\zeta^2}{4} \left\|\tilde{u}\right\|^2 \leq \left\|\Pi_u\right\|^2 = \Pi_u^\top \Pi_u. \tag{38}$$

Multiplying by -1 yields

$$-\Pi_u^\top \Pi_u \leq -\frac{\zeta^2}{4} \left\|\tilde{u}\right\|^2 \tag{39}$$

such that $-\Pi_u^\top \Pi_{uu} \Pi_u$ can now be upper bounded in (33) by

$$\begin{aligned}\dot{V}_1 &\leq -\zeta \Pi_u^\top \Pi_u + \Pi_u^\top (\Pi_{ux}\dot{x} + \Pi_{u\hat{\theta}}\dot{\hat{\theta}} + \Pi_{uw}\dot{w} + \Pi_{ut}) \\ &\leq -\alpha \frac{\zeta^3}{4} \left\|\tilde{u}\right\|^2 + \Pi_u^\top (\Pi_{ux}\dot{x} + \Pi_{u\hat{\theta}}\dot{\hat{\theta}} + \Pi_{uw}\dot{w} + \Pi_{ut})\end{aligned} \tag{40}$$

Let $\|\theta\| \leq c$, therefore it holds that

$$\Pi_u^\top \Pi_{ux}\dot{x} \leq \left\|\Pi_u^\top \Pi_{ux}\right\| (\|f + gu\| + \|\Psi\| c) \tag{41}$$

If α is now chosen as

$$\begin{aligned}\alpha \geq &\frac{4}{\varepsilon_1^2 \zeta^3} \Big((\Pi_u^\top (\Pi_{u\hat{\theta}}\dot{\hat{\theta}} + \Pi_{uw}\dot{w} + \Pi_{ut}) + \\ &\left\|\Pi_u^\top \Pi_{ux}\right\| (\|f + gu\| + \|\Psi\| c))^2 + \frac{1}{4}\Big),\end{aligned} \tag{42}$$

then

$$\begin{aligned}\dot{V}_1 \leq &\Bigg(1 - \frac{\left\|\tilde{u}\right\|^2}{\varepsilon_1^2} \Big(\Pi_u^\top (\Pi_{u\hat{\theta}}\dot{\hat{\theta}} + \Pi_{uw}\dot{w} + \Pi_{ut}) + \left\|\Pi_u^\top \Pi_{ux}\right\| \\ &(\|f + gu\| + \|\Psi\| c)\Big)\Bigg) \Big(\Pi_u^\top (\Pi_{u\hat{\theta}}\dot{\hat{\theta}} + \Pi_{uw}\dot{w} + \Pi_{ut}) + \\ &\left\|\Pi_u^\top \Pi_{ux}\right\| (\|f + gu\| + \|\Psi\| c)\Big) - \frac{1}{4}\frac{\left\|\tilde{u}\right\|^2}{\varepsilon_1^2}.\end{aligned} \tag{43}$$

Independent of the sign of $\Pi_u^\top (\Pi_{ux}\dot{x} + \Pi_{u\hat{\theta}}\dot{\hat{\theta}} + \Pi_{uw}\dot{w} + \Pi_{ut})$, if $\left\|\tilde{u}\right\|^2 > \varepsilon_1^2$ holds then $\dot{V}_1 < 0$.

Analogously, the negativity of $\dot{V}_2$ is shown. It holds that

$$\dot{V}_2 = -2\beta \Upsilon_w^\top \Upsilon_{ww} \Upsilon_w + \Upsilon_{wt}. \tag{44}$$

Using assumption 7, it holds that

$$\dot{V}_2 \leq -\beta \frac{\delta_1^3}{2} \left\|\tilde{w}\right\|^2 + \Upsilon_{wt} \tag{45}$$

If β satisfies

$$\beta \geq \frac{2}{\delta_1^3 \varepsilon_2^2} \left(\Upsilon_{wt}^2 + \frac{1}{4}\right), \tag{46}$$

then

$$\dot{V}_2 \leq -\frac{\left\|\tilde{w}\right\|}{\varepsilon_2^2} \Upsilon_{wt}^2 + \Upsilon_{wt} - \frac{1}{4}\frac{\left\|\tilde{w}\right\|}{\varepsilon_2^2}. \tag{47}$$

Again, independent of the sign of Υ_{wt}, if $\left\|\tilde{w}\right\|^2 > \varepsilon_2^2$ holds then $\dot{V}_2 < 0$

To show the actual convergence of u and w to the vicinities $\|u(t) - u^*(t)\|$ and $\|w(t) - w^*(t)\|$, define $\tilde{\Pi}_u := \Pi_u - \Pi_u^*$ and $\tilde{\Upsilon}_w := \Upsilon_w - \Upsilon_w^*$, where $\Pi_u^* := \Pi_u(x, u^*, w, t)$ and $\Upsilon_w^* := \Upsilon_w(x, u, w^*, t)$. Now introduce the following sets

$$\mathcal{D} := \{\Pi_u \in \mathbb{R}^m, \Upsilon_w \in \mathbb{R}^d : V(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \leq V(\tilde{\Pi}_{u_0}, \tilde{\Upsilon}_{w_0})\}, \tag{48}$$

$$\mathcal{B} := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \mathcal{D} : \|\tilde{u}\| \leq \varepsilon_1 \text{ and } \|\tilde{w}\| \leq \varepsilon_2\}, \tag{49}$$

$$\partial\mathcal{B} := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \mathcal{D} : \|\tilde{u}\| = \varepsilon_1 \text{ and } \|\tilde{w}\| = \varepsilon_2\}, \tag{50}$$

$$\Omega := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \mathcal{D} : \dot{V}_0 \leq 0\}. \tag{51}$$

In the set $\mathcal{B}$, both $\|\tilde{u}\| \leq \varepsilon_1$ and $\|\tilde{w}\| \leq \varepsilon_1$ hold. In the following sets, one of these two conditions does not hold.

$$\mathcal{S}_u := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \mathcal{D} : \|\tilde{u}\| \leq \varepsilon_1 \text{ and } \|\tilde{w}\| \geq \varepsilon_2\}, \tag{52}$$

$$\mathcal{S}_w := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \mathcal{D} : \|\tilde{u}\| \geq \varepsilon_1 \text{ and } \|\tilde{w}\| \leq \varepsilon_2\}. \tag{53}$$

Define $\mathcal{S} := \mathcal{B} \cup \mathcal{S}_u \cup \mathcal{S}_w$. Define the subsets of Ω:

$$\Omega_u := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in (\mathcal{S}_u \cup \mathcal{B}) : \dot{V}_0 \leq 0\}, \tag{54}$$

$$\Omega_w := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in (\mathcal{S}_w \cup \mathcal{B}) : \dot{V}_0 \leq 0\}. \tag{55}$$

Furthermore, define the following sets:

$$\mathcal{E} := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \Omega : \dot{V}_0 = 0\}, \tag{56}$$

$$\mathcal{E}_u := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \Omega_u : \dot{V}_1 = 0\}, \tag{57}$$

$$\mathcal{E}_w := \{(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \Omega_w : \dot{V}_2 = 0\}. \tag{58}$$

If $(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \Omega$, then by the Lyapunov theory (Khalil and Grizzle, 2002), it follows that $(\tilde{\Pi}_u, \tilde{\Upsilon}_w)$ converge to the set $\mathcal{E}$. Since $\dot{V}_0 < 0$ holds for all $(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \Omega \setminus \mathcal{S}$, one has $\mathcal{E} \subseteq \mathcal{S}$. Now, two different cases need to be considered. In the first case, let $(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \Omega_u$. Since $\dot{V}_1 < 0$, $\tilde{\Pi}_u$ converges to the set $\mathcal{E}_u$ with $\mathcal{E}_u \subseteq \Omega_u \subseteq (\mathcal{S}_u \cup \mathcal{B})$. In the second case, $(\tilde{\Pi}_u, \tilde{\Upsilon}_w) \in \Omega_w$ holds. Since $\dot{V}_2 < 0$, $\tilde{\Upsilon}_w$ converges to the set $\mathcal{E}_w$ with $\mathcal{E}_w \subseteq \Omega_w \subseteq (\mathcal{S}_w \cup \mathcal{B})$.

The resulting set $\mathcal{E}_B$ is defined as $\mathcal{E}_B \subseteq \mathcal{E} \cap \mathcal{E}_u \cap \mathcal{E}_w$. Using the fact that $\mathcal{S}_u \cap \mathcal{S}_w = \partial\mathcal{B}$ and $\partial\mathcal{B} \cup \mathcal{B} = \mathcal{B}$, one obtains the following relation

$$\begin{aligned}\mathcal{E}_B &\subseteq \mathcal{E} \cap \mathcal{E}_u \cap \mathcal{E}_w \subseteq \mathcal{S} \cap (\mathcal{S}_u \cup \mathcal{B}) \cap (\mathcal{S}_w \cup \mathcal{B}) \\ &\subseteq (\mathcal{B} \cup \mathcal{S}_u \cup \mathcal{S}_w) \cap (\mathcal{S}_u \cup \mathcal{B}) \cap (\mathcal{S}_w \cup \mathcal{B}) \subseteq \mathcal{B}.\end{aligned} \tag{59}$$

It follows that $(\tilde{\Pi}_u, \tilde{\Upsilon}_w)$ converge to the prescribed vicinities $\mathcal{B}$, which completes the proof. ■

5. CASE STUDY

Consider the following uncertain two-dimensional system.

$$\begin{bmatrix}\dot{x}_1 \\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}x_1^2\theta + x_2 \\ u\end{bmatrix} \tag{60}$$

Here, θ is an unknown parameter. An ACLF for (60) with a corresponding control is

$$\begin{aligned}W(x, \theta) &= \frac{1}{2}x_1^2 + \frac{1}{2}(x_1 + x_2 + x_1^2\theta)^2 \\ v &= -x_1 - (x_1 + x_2 + x_1^2\theta) - (1 - 2x_1\theta)\dot{x}_1\end{aligned} \tag{61}$$

Define

$$\begin{aligned}\xi_1 &:= x_1 \\ \xi_2 &:= x_1 + x_2 + x_1^2\hat{\theta}.\end{aligned} \tag{62}$$

Notice that the origin of (60) is stabilized by a policy v, if the origin of the system

$$\begin{bmatrix}\dot{\xi}_1 \\ \dot{\xi}_2\end{bmatrix} = \begin{bmatrix}-\xi_1 + \xi_2 + \xi_1^2\theta - \xi_1^2\hat{\theta} \\ -\xi_1 + \xi_2 + \xi_1^2\theta + u\end{bmatrix} \tag{63}$$

is stabilized by a policy v. A nominal stabilizing control policy can be designed using backstepping (Krstić et al., 1995) as follows

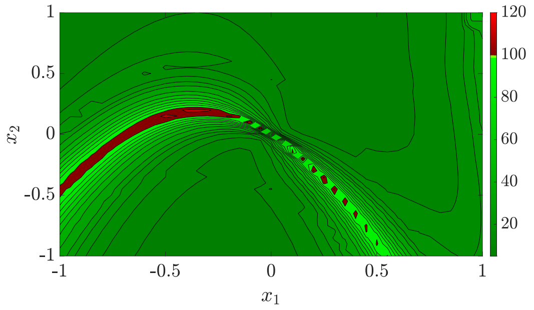

Fig. 1. Relative performance mark (69) of the suggested approach against a nominal stabilizing policy.

$$
\begin{aligned}
u := &-\xi_1 - 100\xi_2 - (\xi_2 - \xi_1)(1 + 2\xi_1\hat{\theta}) \\
&- \gamma\xi_1^4(\xi_1 + \xi_2(1 + 2\xi_1\hat{\theta})).
\end{aligned} \tag{64}
$$

For the system (63) and the control (64), the following ACLF candidate is defined

$$
V(\xi_1, \xi_2, \tilde{\theta}) := \frac{1}{2}\xi_1^2 + \frac{1}{2}\xi_2^2 + \frac{1}{2}\tilde{\theta}^2 \tag{65}
$$

Let the parameter adaptation rule be defined as follows

$$
\dot{\hat{\theta}} := \gamma\xi_1^2(\xi_1 + \xi_2 + 2\xi_1\xi_2\hat{\theta}) \tag{66}
$$

The derivative of the ACLF candidate reads as

$$
\dot{V} = -\xi_1^2 - \xi_2^2. \tag{67}
$$

Let the reward of the consider infinite horizon problem be

$$
r(x, u) := x^\top x + 0.25u^2. \tag{68}
$$

Set $\nu := 0.1(\xi_1^2 + \xi_2^2)$ and $\gamma := 0.001$. Choose the barrier function as $B(z) = -z^{-1}, z \in \mathbb{R}^-$. Further, set $\varepsilon_1 := 0.2$ and $\varepsilon_2 := 0.2$. Choose $\dot{\mu} = -\mu, \mu(0) := 10$. The window of the cost function is chosen as $T := 0.1$. The nominal controller (64) is used to calculate the initial feasible control action. The actor-critic-generated policy is applied to the system and compared to the nominal policy. The performance mark is chosen as

$$
C(x(0), u) = \int_0^{t_{\text{end}}} r(x, u)\mathrm{d}\tau, \tag{69}
$$

where t_{end} is set such that the states reach a vicinity of the equilibrium. To give an accurate measure of the difference in performance, the two control policies are tested for a grid of initial value of the states x_1 and x_2 in the range of $[-1, 1]$. For comparison reasons, Figure 1 shows the relative performance mark, i. e., the fraction of the individual performance marks for both the suggested approach and the nominal policy. The relative performance mark is calculated as

$$
C_{\text{rel}} := \frac{C_{\text{ADP}}(x, u_{\text{ADP}})}{C_{\text{nom}}(x, u_{\text{nom}})} \tag{70}
$$

Areas where the cost of the presented method are smaller than the cost of the nominal controller are given in different shades of green, areas where the nominal control is better are given in shades of red.

6. CONCLUSIONS

This work was concerned with the design of an actor-critic control approach to system with partially known dynamics model. The focus was set to deriving constraints for the control scheme so as to guarantee closed-loop stability of the system's equilibrium. This was achieved by utilizing an adaptive control Lyapunov function. The case study demonstrated cost reduction by the suggested control policy compared to a nominal stabilizing one over a wide range of initial conditions.

REFERENCES

Balakrishnan, S.N., Ding, J., and Lewis, F.L. (2008). Issues on stability of ADP feedback controllers for dynamical systems. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 38(4), 913–918.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, 1st edition.

Bertsekas, D.P. and Tsitsiklis, J.N. (1995). Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, volume 1, 560–564. IEEE.

Fazlyab, M., Paternain, S., Preciado, V.M., and Ribeiro, A. (2016). Interior point method for dynamic constrained optimization in continuous time. In *2016 American Control Conference (ACC)*, 5612–5618. IEEE.

Göhrt, T., Osinenko, P., and Streif, S. (2019a). Adaptive actor-critic structure for parametrized controllers. In *11th IFAC Symposium on Nonlinear Control Systems*.

Göhrt, T., Osinenko, P., and Streif, S. (2019b). Adaptive dynamic programming using lyapunov function constraints. *IEEE Control Systems Letters*, 3(4), 901–906.

Khalil, H.K. and Grizzle, J. (2002). *Nonlinear systems*. Prentice hall Upper Saddle River.

Krstić, M., Kanellakopoulos, I., and Kokotović, P. (1995). *Nonlinear and adaptive control design*. John Wiley & Sons, Inc.

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

Lewis, F.L., Vrabie, D., and Syrmos, V.L. (2012). *Optimal Control*, volume 553. John Wiley & Sons, 3rd edition.

Liu, D. and Wei, Q. (2014). Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 25(3), 621–634.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F.E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234, 11–26.

Nakakuki, T., Shen, T., and Tamura, K. (2008). Adaptive control approach to uncertain longitudinal tire slip in traction control of vehicles. *Asian Journal of Control*, 10(1), 67–73.

Nesterov, Y. (2013). Introductory lectures on convex programming volume i: Basic course. *Springer Science & Business Media*, 87.

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.

Wei, Q., Song, R., and Yan, P. (2016). Data-driven zero-sum neuro-optimal control for a class of continuous-time unknown nonlinear systems with disturbance using adp. *IEEE Transactions on Neural Networks and Learning Systems*, 27(2), 444–458.