Online policy iterations for optimal control of input-saturated systems

Baldi, Simone; Valmorbida, Giorgio; Papachristodoulou, Antonis; Kosmatopoulos, Elias B.

DOI
10.1109/ACC.2016.7526568

Publication date
2016

Document Version
Accepted author manuscript

Published in
Proceedings of the 2016 American Control Conference (ACC 2016)

Accepted Author Manuscript. Link to published article (IEEE): http://dx.doi.org/10.1109/ACC.2016.7526568

Online Policy Iterations for Optimal Control of Input-Saturated Systems

Simone Baldi[1], Giorgio Valmorbida[2], Antonis Papachristodoulou[2] and Elias B. Kosmatopoulos[3]

***Abstract*— This work proposes an online policy iteration procedure for the synthesis of sub-optimal control laws for uncertain Linear Time Invariant (LTI) Asymptotically Null-Controllable with Bounded Inputs (ANCBI) systems. The proposed policy iteration method relies on: a policy evaluation step with a piecewise quadratic Lyapunov function in both the state and the deadzone functions of the input signals; a policy improvement step which guarantees at the same time close to optimality (exploitation) and persistence of excitation (exploration). The proposed approach guarantees convergence of the trajectory to a neighborhood around the origin. Besides, the trajectories can be made arbitrarily close to the optimal one provided that the rate at which the the value function and the control policy are updated is fast enough. The solution to the inequalities required to hold at each policy evaluation step can be efficiently implemented with semidefinite programming (SDP) solvers. A numerical example illustrates the results.**

I. Introduction

This work proposes an online policy iteration procedure for the synthesis of sub-optimal and practically stabilizing control policies for uncertain Linear Time Invariant (LTI) Asymptotically Null-Controllable with Bounded Inputs (ANCBI) systems. This class includes systems with eigenvalues on the imaginary axis (possibly repeated) but no pole with positive real part. The proposed policy iteration is appropriately modified so as to take into account the input saturation function: in particular, the policy evaluation step exploits a class of piecewise quadratic Lyapunov functions which is non-differentiable, but continuous, and depending both on the state and the deadzone function. The policy improvement step is based on a piecewise control policy: the solution of the policy improvement step requires the evaluation of the estimate of the derivative of the Lyapunov function under different candidate control laws, and the resulting mechanism guarantees at the same time close to optimality (exploitation) and persistence of excitation (exploration). The proposed approach guarantees convergence of the trajectory to a neighborhood of the origin. The solution to the inequalities which are required to hold at each step of the policy evaluation is obtained with the solution to semidefinite programmes (SDP).

Synthesis of globally stabilizing control laws for linear saturating systems is a nontrivial problem: even for Linear Time Invariant (LTI) Asymptotically Null-controllable with Bounded Inputs (ANCBI) systems it has been demonstrated with simple examples, that such a class can not, in general, be stabilized by static linear feedback [1]. Different methods to compute globally asymptotically stabilizing nonlinear control laws for ANCBI systems have been proposed [2], [3]. While global stability may not be achieved with linear control laws, strategies for semi-global (exponential) stabilization were presented in [4] (see also the semi-global results for exponentially unstable plants in [5]). However, semi-global results rely on low-gain strategies that may lead to poor performance (in terms of closed-loop convergence rate). In order to obtain faster transients, scheduled [6] and nonlinear control laws [7] have also been proposed in the context of semi-global stabilization. However optimality with respect to criteria other than the convergence rate, has not been explored. In the aforementioned approaches the plant is assumed to be known and the control synthesis is performed offline. An online extension via predictive techniques is considered in [8].

Online techniques for adaptive control of uncertain input-saturated systems have mainly focused on the problem of guaranteeing global stability [9], [10] without optimality considerations: these schemes guarantee global stability via a continuous-time direct adaptive controller. More recently, approaches to optimal control of input-saturated systems have been developed, with the aim of approximating the optimal solution to the Hamilton-Jacobi-Bellman equation. Since some knowledge of the dynamics is required to implement these techniques, online estimation must be employed. Interesting approaches, which do not take input-saturation into account, are [11], [12], where actor-critic structures are combined with a third network meant to approximate the unknown system dynamics. Actor-critic structures are updated in such a way to approximate the optimal control solution and the optimal value function respectively. For constrained-input systems some offline [13] and online [14], [15], [16] actor-critic methods based on neural networks have been proposed, where however, the input saturation is assumed to be a sigmoidal continuous and differentiable function.

The paper is organized as follows: Section II presents the problem formulation; Section III recalls an offline policy iteration approach for input saturated systems; Section IV

[1]S. Baldi is with the Delft Center for Systems and Control, Delft University of Technology, Delft 2628CD, The Netherlands s.baldi@tudelft.nl

[2]G. Valmorbida and A. Papachristodoulou are with Department of Engineering Science, Control Group, University of Oxford, Parks Road, Oxford OX1 3PJ, U.K. G. Valmorbida is also affiliated to Somerville College, University of Oxford, Oxford, U.K. A. Papachristodoulou was supported in part by the Engineering and Physical Sciences Research Council projects EP/J012041/1, EP/I031944/1 and EP/J010537/1. giorgio.valmorbida@eng.ox.ac.uk, antonis@eng.ox.ac.uk

[3]E. B. Kosmatopoulos is with Dept. of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi 67100, Greece and Informatics & Telematics Institute, Center for Research and Technology Hellas (ITI-CERTH), Thessaloniki 57001, Greece kosmatop@iti.gr

presents the estimation scheme for the uncertain system dynamics; Section V contains the online policy iteration approach for input-saturated systems, and Section VI the numerical implementation of the policy evaluation step. The numerical example in Section VII demonstrates the effectiveness of the proposed approach.

II. Problem Formulation

We study the class of uncertain LTI Asymptotically Null-Controllable with Bounded Inputs (ANCBI) systems in the presence of input saturation, which consists of the set of dynamic linear systems without exponentially unstable modes. Consider the input-saturated system

$$\dot{x} = A(\Theta^*)x + B(\Theta^*)sat\left(u(x)\right), \quad x(0) = x_0, \tag{1}$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $A(\Theta^*) \in \mathbb{R}^{n\times n}$ and $B(\Theta^*) \in \mathbb{R}^{n\times m}$, $\max(\Re(\lambda(A))) \leq 0$. Both A and B are assumed to be matrices with unknown entries represented by Θ^*. The function $sat : \mathbb{R}^m \to \mathscr{U} \subset \mathbb{R}^m$ is a vector saturation function, with entries satisfying

$$\left(sat\left(u(x)\right)\right)_j = \begin{cases} \overline{u}_j, & \text{if} \quad u_j > \overline{u}_j \\ u_j, & \text{if} \quad \underline{u}_j \leq u_j \leq \overline{u}_j \\ \underline{u}_j, & \text{if} \quad \underline{u}_j > u_j \end{cases}$$

with $\overline{u}_j$ and $\underline{u}_j$ the upper and lower bound of the j-th input, respectively. The set of admissible inputs is defined as

$$\mathscr{U} := \left\{u \in \mathbb{R}^m | \underline{u}_j \leq u_j \leq \overline{u}_j, j = 1,\ldots,m\right\}.$$

In the following, for convenience of notation, we introduce the dead-zone function $dz(u(x)) := u(x) - sat\left(u(x)\right)$, and rewrite (1) as

$$\dot{x} = A(\Theta^*)x + B(\Theta^*)u(x) - B(\Theta^*)dz\left(u(x)\right), \quad x(0) = x_0. \tag{2}$$

We also introduce a cost function for the system (2) in the form

$$J = \int_0^\infty L(x,u)dt = \int_0^\infty \left[x'Qx + \left(sat\left(u\right)\right)'Rsat\left(u\right)\right]dt, \tag{3}$$

where the prime denotes transpose. To address the parametric uncertainty in the system, we will develop an adaptive control policy, combined with a parametric adaptation law taking the following form

$$\dot{\hat{\Theta}}(t) = p(\hat{\Theta}(t), \Xi(t)), \quad \hat{\Theta}(0) = \Theta_0 \tag{4a}$$

$$\dot{\Xi}(t) = g(\Xi(t), x(t), u(t)), \quad \xi(0) = \xi_0 \tag{4b}$$

$$V(t) = s(\hat{\Theta}(t), u(t)), \tag{4c}$$

$$u^+(t) = h(\hat{\Theta}(t), V(t)), \tag{4d}$$

where $\hat{\Theta}$ are the estimates of Θ^*, Ξ are auxiliary variables used for estimation, V indicates the value function, and $u^+(t)$ indicates the feedback law to be used in the time interval $[t+k\delta t, t+(k+1)\delta t]$, where δt is the sampling time. The mappings, *p, g, s, h* will be designed to guarantee the convergence of the state in a neighborhood of the origin and to optimize the cost (3).

Let us introduce the definitions below:

Definition 1: [Practical stability [17]] Given a nonlinear system $\dot{x} = f(x)$, with $f(0) = 0$, the origin of the system is *practically stable* if, for given $(c, \overline{c})$ with $0 < c < \overline{c}$, every solution $x(t, x_0)$ of the system satisfies

$$\|x_0\| < c \Rightarrow \|x(t, x_0)\| < \overline{c}, \quad t \geq t_0$$

for some $t_0 \in \mathbb{R}_+$.

Definition 2: [Asymptotic minimization] Given a function $J(\vartheta)$ and

$$\vartheta^* = \arg\min_{\vartheta} J(\vartheta)$$

the sequence $\{\vartheta_k\}$ asymptotically minimizes J if $\lim_{k\to\infty} \vartheta_k = \vartheta^*$.

The objective of the control problem can be stated as:

Problem 1: Design the functions $p(\cdot,\cdot)$, $g(\cdot,\cdot,\cdot)$, $s(\cdot,\cdot)$ and $h(\cdot,\cdot)$ so that the closed-loop (2)-(4) guarantees the practical stability of the origin of (2) and the asymptotic minimization of the cost (3).

In the following, multidimensional vectors are intended as column vectors, while the gradient of a scalar quantity with respect to a vector is intended as a row vector. We introduce the sector condition pertaining to the deadzone presented in [18]. The deadzone function $dz(u(x))$ satisfies the following sector inequality

$$dz'(u(x))\Pi_1\left(u(x) - dz(u(x))\right) \geq 0, \quad \forall\, x \in \mathbb{R}^n. \tag{5}$$

implying that the deadzone function is contained in the sector $[0, I]$. Furthermore, define $\phi(x) := \frac{d\, dz(u(x))}{dt}$ satisfying

$$\phi(x) = \begin{cases} 0 & \text{if } dz\left(u(x)\right) = 0 \\ \dot{u}(x) & \text{if } dz\left(u(x)\right) \neq 0, \end{cases} \tag{6}$$

which can be expressed by the two equalities

$$\phi'(x)\Pi_2\left(\dot{u}(x) - \phi(x)\right) = 0 \tag{7}$$

$$dz'\left(u(x)\right)\Pi_3\left(\dot{u}(x) - \phi(x)\right) = 0, \tag{8}$$

where $\Pi_1, \Pi_2, \Pi_3 \in \mathbb{R}^{m\times m}_{diag}$ are diagonal matrices, and Π_1 is positive definite. Due to the monotonicity of the saturation and the deadzone functions, we also have that the following inequality holds for two arbitrary control policies $u(x)$ and $v(x)$

$$\left(dz\left(u(x)\right) - dz\left(v(x)\right)\right)'\left(sat\left(u(x)\right) - sat\left(v(x)\right)\right) > 0. \tag{9}$$

We adopt the well-known result from optimal control theory [19, Chap.3], that states that the optimal control policy $u^o(x)$ that minimizes (3) satisfies

$$u^o = \arg\min_{u(\cdot)\in\mathscr{U}} \left\{\frac{d\,V^o}{dx}\left(Ax + Bu\right) + L(x,u)\right\}, \tag{10}$$

where $V^o(x)$ is the value function (or cost-to-go function) that solves the Hamilton-Jacobi-Bellman (HJB) equation

$$\min_{u(\cdot)\in\mathscr{U}} \left\{\frac{d\,V^o}{dx}\left(Ax + Bu\right) + L(x,u)\right\} = 0. \tag{11}$$

In order to have a well-posed problem we make the following assumption

Assumption 1: There exists a globally stabilizing control policy $\overline{u}$ for system (1).

According to standard converse-Lyapunov results [20], Assumption 1 implies the existence of a continuous, positive definite, radially unbounded control Lyapunov function (CLF) $\bar{V}:\mathbb{R}^n\to\mathbb{R}_+$ which satisfies

$$\min_{u(\cdot)\in\mathscr{U}}\left\{\frac{d\bar{V}}{dx}(Ax+Bu)\right\}<0,\quad \forall x\neq 0.$$

The following lemma relates the CLF to the uncontrollable region of system (1):

Lemma 1: Assumption 1 implies that there exist positive constants ε_i, $i=1,2,3$ *such that the following condition holds, for all* $x\in\mathbb{R}^n$,

$$\left|\frac{d\bar{V}}{dx}(x)B\right|<\varepsilon_1 \text{ and } |x|>\varepsilon_3 \Rightarrow \frac{d\bar{V}}{dx}(x)Ax<-\varepsilon_2. \quad (12)$$

Let us define the *uncontrollable region* of (1) to be the subset $\mathscr{R}$ defined according to

$$\mathscr{R}=\left\{x\in\mathbb{R}^n:|x|>\varepsilon_3 \text{ and } \left|\frac{\partial\bar{V}}{\partial x}(x)B\right|<\varepsilon_1\right\}.$$

Note that condition (12) implies that for $x\in\mathscr{R}$, the choice $u=0$ guarantees that $\dot{V}<0$.

III. Offline Policy Iterations under Saturation Constraints

The iterative strategy in Algorithm 1 was presented in [21] and provides an offline solution to Problem 1.

Algorithm 1 Modified policy iteration

1: *Initialize*:
2: $c\leftarrow 0$.
3: $\bar{u}^c_{pw}\leftarrow u^0$.
4: $u^c_{pw}\leftarrow u^0$.
5: *Policy evaluation*:
6: Given u^c_{pw}, **solve** for $V^c(x)=W^c(x,dz(u^c(x)))$
7:
$$\frac{dV^c(x)}{dx}\left(Ax+Bsat\left(u^c_{pw}\right)\right)+L(x,u^c_{pw})=0 \quad (13)$$
8: *Feasibility*:
9: With $W^c(x,dz(u^c(x)))$ of *Policy evaluation*, **check**
10:
$$\frac{dV^c(x)}{dx}\left(Ax+Bsat\left(u^c\right)\right)+L(x,u^c_{pw})<0 \quad (14)$$
11: **if** (13) is *feasible*, $\bar{u}^c(x)\leftarrow u^c(x)$
12: **else** $\bar{u}^c(x)\leftarrow u^{(c-1)}(x)$
13: *Policy improvement*:
14: **Update** the piecewise control policy
15:
$$u^{c+1}_{pw}=\begin{cases}-\frac{1}{2}R^{-1}B'\left.\frac{\partial W^{c\prime}}{\partial x}\right|_{q^c=0} & \text{in } \Xi^c_1\cup\Xi^c_2\cup\Xi^c_3\\ \bar{u}^c & \text{in } \Xi^c_4\end{cases} \quad (15)$$
16:
17: **if** $\Delta W^c(x(0)):=W^c(x(0))-W^{(c-1)}(x(0))<\delta$, **STOP**
18: **else** $c\leftarrow c+1$, **goto** *Policy improvement*.

In the algorithm, the policy evaluation and the policy iteration steps are performed based on the piecewise value function and piecewise control policy defined as follows. Given the value function $V^c(x)=W^c(x,dz(u^c))$ that solves (13), define the following *approximated* policy improvement

$$u^{c+1}_{ap}(x) = -\frac{1}{2}R^{-1}B'\left.\frac{\partial W^{c\prime}}{\partial x}\right|_{dz(u^c)=0}, \quad (16)$$

In order to discuss the properties of the policy (16) let us define the following state-space partition arising from

$$\Omega^c(x)=\{x:dz(u^c(x))=0\} \quad (17)$$
$$\Omega^{c+1}(x)=\left\{x:dz(u^{c+1}_{ap}(x))=0\right\}, \quad (18)$$

$$\begin{array}{lll}\Xi^c_1 := \Omega^c\cap\Omega^{c+1} & \text{(Region 1)}\\ \Xi^c_2 := \Omega^c\backslash\Omega^{c+1} & \text{(Region 2)}\\ \Xi^c_3 := \Omega^{c+1}\backslash\Omega^c & \text{(Region 3)}\\ \Xi^c_4 := \mathbb{R}^n\backslash(\Omega^c\cup\Omega^{c+1}) & \text{(Region 4)}\end{array}$$

and satisfying $\cup_i\Xi^c_i=\mathbb{R}^n$ and $\Xi^c_i\cap\Xi^c_j=\emptyset$, $i\neq j$.

To study the stability properties of the policy u^{c+1}_{ap}, given a globally stabilizing policy u^c and a value function W^c that certify global stability, we define the piecewise policy

$$sat\left(u^{c+1}_{pw}\right)=\begin{cases}sat\left(u^{c+1}_{ap}\right) & \text{in } \Xi^c_1\cup\Xi^c_2\cup\Xi^c_3\\ sat\left(u^c\right) & \text{in } \Xi^c_4\end{cases} \quad (19)$$

with the value function

$$W^c_{pw}:=\begin{cases}W^c_{un} & \text{in } \Xi^c_1\cup\Xi^c_2\cup\Xi^c_3\\ W^c & \text{in } \Xi^c_4\end{cases} \quad (20)$$

where $W^c_{un}\in\mathscr{C}^1$ is the unsaturated value function defined as

$$W^c_{un}(x):=W^c(x,0). \quad (21)$$

We obtain the following result

Proposition 1: The piecewise value function (20) certifies the global stability of the piecewise control policy (19).

Proof: See [21]. ■

IV. Estimation of the System Dynamics

The results of Section III require the knowledge of matrices A and B. Its extension to the uncertain system (1) requires an online parameter estimator. This task will be performed with standard techniques for parameter estimation. To this purpose we write (1) as

$$\dot{x}=A_mx+(A-A_m)x+Bsat(u), \quad (22)$$

with A_m a Hurwitz matrix. We use the series-parallel parametric model [22] to obtain

$$\dot{\hat{x}}=A_m\hat{x}+(\hat{A}-A_m)x+\hat{B}sat(u), \quad (23)$$

where $\hat{x}$ is the state of the parametric model and $\hat{A}$, $\hat{B}$ are the matrices to be estimated. In order to develop a linear-in-the-parameters model for (22) we filter every component of $\dot{x}$, x and $sat(u)$ with a stable filter $\lambda/(s+\lambda)$, $\lambda>0$

$$z_f=\frac{s\lambda}{s+\lambda}x, \quad (24a)$$

$$x_f=\frac{\lambda}{s+\lambda}x, \quad (24b)$$

$$\nu_f = \frac{\lambda}{s+\lambda} sat(u). \tag{24c}$$

We thus obtain

$$z_f = A_m x_f + (A - A_m) x_f + B\nu_f, \tag{25}$$

and similarly for (23)

$$\hat{z}_f = A_m \hat{x}_f + (\hat{A} - A_m) x_f + \hat{B}\nu_f, \tag{26}$$

where z_f, x_f, and ν_f are all measurable signals to be used for the estimator. After collecting all the entries of A and B in $\Theta^* = [A\ B]'$ and defining $\hat{\Theta} = \left[\hat{A}\ \hat{B}\right]'$[1], we adopt a parameter estimator based on integral cost and gradient update [22], so as to obtain

$$\dot{\hat{\Theta}} = \mathsf{P}\left(-\gamma \bar{R}\hat{\Theta} - \gamma \bar{Q}\right), \quad \hat{\Theta}(0) = \Theta_0 \tag{27a}$$

$$\dot{\bar{R}} = -\beta \bar{R} + \left[x_f'\ \ \nu_f'\right]' \left[x_f'\ \ \nu_f'\right], \quad \bar{R}(0) = 0 \tag{27b}$$

$$\dot{\bar{Q}} = -\beta \bar{Q} - \left[x_f'\ \ \nu_f'\right]' z_f', \quad \bar{Q}(0) = 0 \tag{27c}$$

where β and γ are positive constants and P denotes a projection operator which has to be designed to keep the estimates inside a convex set.

The estimation law (27) satisfies the following properties:

Theorem 1: [22]

i) $\varepsilon := z_f - \hat{z}_f \in \mathscr{L}_2 \cap \mathscr{L}_\infty$

ii) $\lim_{t\to\infty} \left|\dot{\hat{\Theta}}\right| = 0$

iii) if $\left[x_f'\ \ \nu_f'\right]'$ is persistently exciting, then $\hat{\Theta} \to \Theta^*$ exponentially and the rate of convergence increases with γ.

V. Online Policy Iterations under saturation constraints

Algorithm 1 is now revised for online implementation. Differently from Section IV, the iterations are not implemented offline at each step c, $c \in \mathbb{Z}_+$, but online at each time instant t_k, $t_k = 0, \Delta t, 2\Delta t, \ldots$, where Δt is the update sample time. The proposed algorithm is shown in Algorithm 2.

In Algorithm 2, t_k^+ indicates the instant of time at which the previous policy is updated, $\left\{\pm u_{(j)}^k, j = 1, \ldots, n\right\}$ indicates a set of candidate policies, $\hat{V}_{(\pm j)}^k(t_k)$ in (34) indicates the estimates of the derivative of the value function calculated at time t_k with the corresponding policy $\pm u_{(j)}^k$. Furthermore, the candidate control policies $\pm u_k^{(j)}$ are calculated as follows

$$h^k(\zeta, x) = -\frac{1}{2} R^{-1} \hat{B}' \left.\frac{\partial V^{k\prime}}{\partial x}\right|_{dz(u^{(k-1)})=0} \tag{36}$$

$$\pm u_{(j)}^k = h^k(\zeta \pm \Delta\zeta_{(j)}, x), \tag{37}$$

[1] A parameter estimator can be developed also in the case where only a subset of entries of A and B needs to be estimated, by bringing to the left-hand side of (25) and (26) all the quantities that are known and do not need to be estimated.

Algorithm 2 Online policy iteration

Initialize:
  $k \leftarrow 0$.
  $\bar{u}_{pw}^k \leftarrow u^0$.
  $u_{pw}^k \leftarrow u^0$.
Policy evaluation:
  Given u_{pw}^k, $\hat{A}^{(k-1)} = \hat{A}(t_{k-1})$, $\hat{B}^{(k-1)} = \hat{B}(t_{k-1})$,
  solve for $V^k(x) = W^k(x, dz\left(u^k(x)\right))$

$$\frac{d\,V^k(x)}{dx}\left(\hat{A}^{(k-1)}x + \hat{B}^{(k-1)} sat\left(u_{pw}^{(k-1)}\right)\right) + L(x, u_{pw}^{(k-1)}) = 0 \tag{28}$$

$$\frac{d\,V^k(x)}{dx}\left(\bar{A}x + \bar{B} sat\left(u_{pw}^{(k-1)}\right)\right) + L(x, u_{pw}^{(k-1)}) < 0 \tag{29}$$

$$\bar{A} = \hat{A}^{(k-1)} + \Delta A, \bar{B} = \hat{B}^{(k-1)} + \Delta B \text{ with } \Delta A, \Delta B \in \mathscr{N}^k \tag{30}$$

$$\mathscr{N}^k = \left\{\Delta A, \Delta B \mid \Delta A'\Delta A \prec \eta_1^k I,\ \Delta B'\Delta B \prec \eta_2^k I\right\}, \tag{31}$$

Feasibility:
  With $W^k(x, dz\left(u^k(x)\right))$ of *Policy evaluation*, **check**

$$\frac{d\,V^k(x)}{dx}\left(\hat{A}^{(k-1)}x + \hat{B}^{(k-1)} sat\left(u^k\right)\right) + L(x, u_{pw}^k) < 0 \tag{32}$$

  if (13) is *feasible*, $\bar{u}^k(x) \leftarrow u^k(x)$
  else $\bar{u}^k(x) \leftarrow u^{(k-1)}(x)$
Estimation:
  Update the estimates $\hat{A}(t_k)$, $\hat{B}(t_k)$ according to (27), with $\mathsf{P} = \mathsf{P}_k$ the projection operator that keeps the estimate inside the set $\mathscr{N}^k$.
Policy improvement:
  Update the piecewise control policy

$$u(t_k^+) = \arg \min_{\pm u_{(j)}^k, j=1,\ldots,n} \hat{V}_{(\pm j)}^k(t_k), \tag{33}$$

$$\begin{aligned}\hat{V}_{(\pm j)}^k(t_k) &= \frac{\partial V^k}{\partial x}\left[\hat{A}(t_k)x(t_k) + \hat{B}(t_k)(\pm u_{(j)}^k)\right] \\ &\quad + Q(x(t_k)) + u_{(j)}^k R u_{(j)}^k,\end{aligned} \tag{34}$$

$$u_{pw}^{k+1} = \begin{cases} u(t_k^+) & \text{in } \Xi_1^k \cup \Xi_2^k \cup \Xi_3^k \\ \bar{u}^k & \text{in } \Xi_4^k \end{cases} \tag{35}$$

  if $\Delta W^k(x(0)) := W^k(x(0)) - W^{(k-1)}(x(0)) < \delta$, **STOP** updating W and u
  else goto *Policy evaluation*.
  $k \leftarrow k+1$

where ζ are the coefficients of the expression in (36), and $\Delta\zeta^{(j)}$ are zero-mean random vectors in $[-2\alpha_k,-\alpha_k]^n \cup [\alpha_k,2\alpha_k]^n$ satisfying

$$\left|\left[\Delta\zeta^{(1)},\ldots,\Delta\zeta^{(n)}\right]\right|^{-1} < \frac{\Xi}{\alpha_k}, \tag{38}$$

where α_k is a positive sequence and Ξ is a finite positive number independent of α_k. The following result is given.

Theorem 2: Let Δt be sufficiently small. Then, for arbitrary small $\bar{\alpha} > 0$, there exist finite positive constants β_1, β_2, γ_l and a finite positive integer $\bar{h} = \mathscr{O}\left(\frac{1}{\gamma}\right)$ such that the following condition holds: if α_k satisfies

$$\begin{cases} 0 < \alpha_k \leq \beta_2 & \text{if } \left|\frac{d\,V^k}{dx}\hat{B}\right| < \hat{\varepsilon}_1 \text{ or } k \leq \bar{h} \\ \alpha_k \geq \beta_1 & \text{otherwise} \end{cases}$$

where $\hat{\varepsilon}_1$ is a positive design constant satisfying

$$\frac{1}{4}\varepsilon_1 < \hat{\varepsilon}_1 \leq \frac{1}{2}\varepsilon_1$$

and the adaptive gain γ of the estimator satisfies

$$\gamma \geq \gamma_l$$

then, the proposed adaptive control scheme guarantees that the closed-loop solutions are bounded and, moreover,

$$\limsup_{t\to\infty} |x(t)| \leq \varepsilon_3, \text{ w.p.1}$$

and

$$-\bar{\alpha} < \dot{\mathscr{L}}(t_k^+) < 0, \text{ if } x_k \notin \mathscr{R} \text{ or } (x_k,\hat{\theta}_k) \notin \mathscr{S}_k, \text{ w.p.1}$$

where

$$\dot{\mathscr{L}}(t_k^+) = \min_{u(\cdot)\in\mathscr{U}} \left\{ \frac{d\,V^o}{dx}(Ax+Bu) + L(x,u) \right\}$$

and $\mathscr{S}_k$ is a subset of $\mathbb{R}^n \times \mathbb{R}^{n\times(n+m)}$ that satisfies $\mathscr{S}_k = \emptyset, \forall k > \bar{h}$.

Proof: Following similar lines as in [23] (not shown for lack of space). ■

Remark 1: Each policy evaluation step (28)-(29) returns a set of plants that are stabilized by the control law u_{pw}^k. Such a set is given by $\mathscr{N}^k$ in (31)*. This set is used in the estimation law (27) to project the estimate. This approach resembles the so-called 'certainty equivalence principle' of adaptive control [22], where the control policy is stabilizing for the estimated plant and it is updated by solving the underlying control problem for the estimated plant.*

Remark 2: The rationale behind (33) *is that among a set of possible candidate control laws, the one that minimizes* (34)*, i.e. that more closely satisfies the HJB equation is chosen. This choice guarantees the so-called 'exploitation task' of the control policy. Furthermore, the candidate control policies are generated randomly so as to satisfy condition* (38)*: this guarantees the so-called 'exploration task' of the control policy, i.e. persistence of excitation and convergence of the estimates to their real value. It can be shown that the Bernoulli distribution satisfies condition* (38) *[23]: other distributions (segmented uniform, U-shaped) are also possible [24, Sect. 7.3].*

VI. Numerical example

In the following, we present a numerical example to illustrate the results obtained via the proposed policy iterations. The procedure has been implemented in SOSTOOLS [25] and the formulated SDPs were solved with SeDuMi [26]. The dimension of the example helps to illustrate the results by plotting the computed value functions and the time-evolution of the control policies. It is also worth mentioning that as the number of variables and the degrees of the involved polynomials increase, the dimensions of the related SDPs can be large.

Consider the following 1-input 1-state system

$$\dot{x}(t) = -ax + bsat(u(t)), \quad x(0) = -1 \tag{39}$$

with a and b two positive and unknown constants. The saturation bounds are $-0.5 \leq u \leq 0.5$ and the initial globally stabilizing (but not optimal) state-feedback $u(x) = -0.3x$. For this system we consider the cost as in (3) with $Q = 1$ and $R = 1$. For $a = 1$, $b = 1$, $\hat{a}(0) = 2$, $\hat{b}(0) = 1.5$, $\beta = 3$, $\lambda = 3$, $\Gamma = 10$, $\alpha_k = 0.025$, $\Delta t = 0.01$ we apply the proposed online policy iterations.

The simulation is run for 5 seconds. Fig. 1 shows the online evolution of the state and input with the proposed adaptive law. Finally Fig. 2 shows the offline evolution of the cost using the controllers synthesized at every time step: it can be observed that the cost is monotonically decreasing. The online evolution of the Hamilton-Jacobi-Bellman equation is also shown: it can be observed that for the presented example the controller synthesized at every time step are stabilizing not only the estimated plant, but also the actual plant.

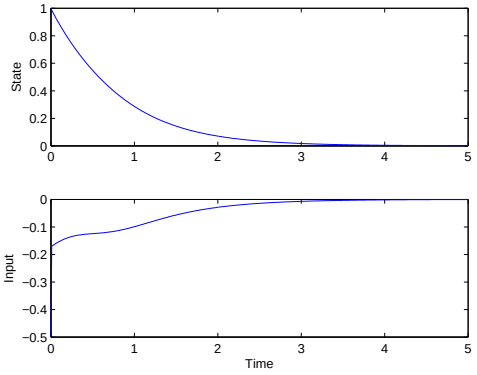

Fig. 1: Online state (upper) and input (lower) evolution.

VII. Conclusions

This work proposed an online policy iteration procedure for the synthesis of approximately optimal control laws for uncertain Linear Time Invariant (LTI) Asymptotically Null-Controllable with Bounded Inputs (ANCBI) systems. The proposed policy iteration method relies on: a policy evaluation step with a piecewise quadratic Lyapunov function which is non-differentiable, but continuous, and polynomial

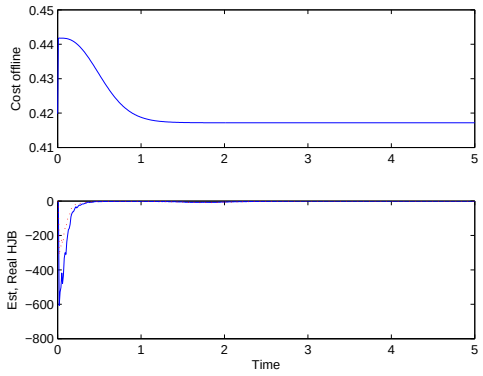

Fig. 2: Offline evolution of the cost using the controllers synthesized at every time step (upper) and online evolution of the Hamilton-Jacobi-Bellman equation (lower). With a solid line is the HJB for the estimated plant, and with a dashed line is the HJB for the actual plant.

in both the state and the deadzone functions of the input signals; a policy improvement step which guarantees at the same time close to optimality (exploitation) and persistence of excitation (exploration). The proposed approach guarantees convergence of the trajectory to a neighborhood around the origin. Besides, the trajectories can be made arbitrarily close to the optimal one provided that the rate at which the the value function and the control policy are updated is fast enough.

Future work includes the extension of the proposed methodology to linear systems with exponentially unstable modes for which only local stability is achievable. Such an extension is under study and will account for generalized sector condition which is instrumental to compute region of attraction estimates. We will also generalize the obtained conditions to systems defined by polynomial vector fields and polynomial input matrices.

REFERENCES

[1] A. Fuller, "In-the-large stability of relay and saturating control systems with linear controllers," *International Journal of Control*, vol. 10, pp. 457–480, 1969.

[2] E. Sontag and H. Sussmann, "Nonlinear output feedback design for linear systems with saturating controls," in *Decision and Control, 1990., Proceedings of the 29th IEEE Conference on*, Dec 1990, pp. 3414–3416 vol.6.

[3] H. Sussmann, E. Sontag, and Y. Yang, "A general result on the stabilization of linear systems using bounded controls," *Automatic Control, IEEE Transactions on*, vol. 39, no. 12, pp. 2411–2425, Dec 1994.

[4] Z. Lin and A. Saberi, "Semi-global exponential stabilization of linear systems subject to "input saturation" via linear feedbacks," *Systems & Control Letters*, vol. 21, no. 3, pp. 225 – 239, 1993.

[5] B. Zhou, G. Duan, and Z. Lin, "Global stabilization of the double integrator system with saturation and delay in the input," *IEEE Transactions on Circuits and Systems-I*, vol. 57, no. 6, pp. 1371–1383, 2010.

[6] D. Henrion, G. Garcia, and S. Tarbouriech, "Piecewise-linear robust control of systems with input constraints," *European Journal of Control*, vol. 5, no. 1, pp. 157–166, 1999.

[7] G. Valmorbida, L. Zaccarian, S. Tarbouriech, I. Queinnec, and A. Papachristodoulou, "A polynomial approach to nonlinear state feedback stabilization of saturated linear systems," in *Decision and Control, 2014, Proceedings of the 53rd IEEE Conference on*, Dec 2014, pp. 6317–6322.

[8] M. Tanaskovic, L. Fagiano, R. Smith, and M. Morari, "Adaptive receding horizon control for constrained mimo systems," *Automatica*, vol. 50, pp. 3019–3029, 2014.

[9] F. Z. Chaoui, F. Giri, J. M. Dion, M. M-Saad, and L. Dugard, "Direct adaptive control subject to input amplitude constraint," *IEEE Transactions on Automatic Control*, vol. 45, pp. 485–490, 2000.

[10] C. Zhang and R. J. Evans, "Continuous direct adaptive control with saturation input constraint," *IEEE Transactions on Automatic Control*, vol. 39, pp. 1718–1722, 1994.

[11] P. J. Werbos, "Approximate dynamic programming for real time control and neural modelling," in *Handbook of Intelligent Control*, D. A. White and D. A. Sofge, Eds. Multiscience Press, Brentwood, U.K., 1992.

[12] S. Bhasin, R. Kamalapurkar, M. Johnson, K. Vamvoudakis, F. L. Lewis, and W. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, pp. 82–92, 2013.

[13] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, pp. 779–791, 2005.

[14] H. Zhang, Y. Luo, and D. Liu, "Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints," *IEEE Tranactions on Neural Network*, vol. 20, pp. 1490–1503, 2009.

[15] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, pp. 1513–1525, 2013.

[16] ——, "Online solution of nonquadratic two-player zero-sum games arising in the h_∞ control of constrained input systems," *International Journal of Adaptive Control and Signal Processing*, vol. 24, pp. 232–254, 2013.

[17] J. Lasalle and S. Lefschetz, *Stability by Liapunov's direct method: with applications*. Academic Press, New York, 1967.

[18] D. Dai, T. Hu, A. R. Teel, and L. Zaccarian, "Piecewise-quadratic lyapunov functions for systems with deadzones or saturations," *Systems & Control Letters*, vol. 58, no. 5, pp. 365 – 371, 2009.

[19] D. E. Kirk, *Optimal Control Theory: An Introduction*. Prentice-Hall, Englewood Cliffs, N.J., 1970.

[20] R. A. Freeman and P. V. Kokotovic, "Inverse optimality in robust stabilization," *SIAM Journal on Control and Optimization*, vol. 34, pp. 1365–1391, 1996.

[21] S. Baldi, G. Valmorbida, A. Papachristodoulou, and E. B. Kosmatopoulos, "Piecewise polynomial policy iterations for synthesis of optimal control laws in input-saturated systems," *Proceedings of the 2015 American Control Conference*, pp. –, july 2015.

[22] P. A. Ioannou and J. Sun, *Robust Adaptive Control*. Dover Publications, 2012.

[23] E. B. Kosmatopoulos, "An adaptive optimization scheme with satisfactory transient performance," *Automatica*, vol. 45, no. 3, pp. 716–723, 2009.

[24] J. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, Hoboken, NJ, 2003.

[25] A. Papachristodoulou, J. Anderson, G. Valmorbida, S. Prajna, P. Seiler, and P. A. Parrilo, *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*, `http://arxiv.org/abs/1310.4716`, 2013, available from `http://www.eng.ox.ac.uk/control/sostools`, `http://www.cds.caltech.edu/sostools` and `http://www.mit.edu/~parrilo/sostools`.

[26] L. Peaucelle and D. Henrion, "Sedumi interface 1.02: a tool for solving lmi problems with sedumi," *Proceedings, IEEE International Symposium on Computer Aided Control System Design*, pp. 272–277, 2002.