Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-5 (2016) 067–072

Improvement of LMI controllers of Takagi-Sugeno models via Q-learning ⋆

Henry Díaz, Leopoldo Armesto, Antonio Sala

Inst. Univ. Autom. Inf. Industrial (AI²), Inst. Diseño Fabric. (IDF), Universitat Politècnica de València, C/Camino de Vera s/n, 46022, Valencia, Spain (e-mail: hendia@posgrado.upv.es).

Abstract: This paper presents a preliminary attempt to bridge the conservative (shape-independent) results from guaranteed-cost LMIs and the reinforcement learning setups which learn optimal controllers from data. In this sense, the proposed approach uses an initialization based on the LMI solution and proposes an approximation of the Q-function using polynomials of the membership functions in Takagi-Sugeno models. The resulting controller is shape-dependent, that is, uses the knowledge of membership functions and data to clearly improve LMI solutions.

Keywords: Reinforcement learning, adaptive dynamic programming, Q-learning, Takagi-Sugeno, LMI.

1. INTRODUCTION

Nonlinear systems can be modeled as the so-called Takagi-Sugeno systems, via the well-known sector-nonlinearity approach (Tanaka and Wang, 2001). Usual developments for Takagi-Sugeno systems regarding optimal control set up LMI conditions in order to find the so-called guaranteed-cost Lyapunov solutions (Wu and Cai, 2004). Such solutions are shape-independent (i.e., the actual membership function shape is assumed to be known at runtime, but not at design time) and, hence, conservative (Sala, 2009).

On the other hand, there is plenty of literature related to Reinforcement learning, e.g. (Sutton and Barto, 1998), neuro-dynamic programming (Bertsekas and Tsitsiklis, 1996), approximate dynamic programming (Powell, 2011), adaptive dynamic programming (Zhang et al., 2012), Q-learning (Watkins and Dayan, 1992), etc. which dates from Bellman's optimality principle in the 1950s with viable computational implementations starting in the late 1980s. Reinforcement learning approaches have been mathematically formalised in a control related context with the aim to convert these ideas into practically feasible approaches e.g.(Lewis and Vrabie, 2009), (Lewis et al., 2012), (Lewis and Liu, 2013) (Kiumarsi et al., 2014).

The objective of this paper is bridging these two approaches to optimal control of nonlinear systems: given a nonlinear system in Takagi-Sugeno form, the LMI solution and the PDC-like controller structures associated to them will inspire a particular parametrization of Q-learning algorithms so that such learning algorithm can be initialized with the LMI solution. Also, higher-dimensional summations will allow improving the learning solution.

⋆ The authors are grateful to projects DPI2011-27845-C02-01 and DPI2013-42302-R from Spanish Government, Grant PROMETEOII/2013/004 from Generalitat Valenciana and Ph.D. grant SENESCYT from the Government of Ecuador.

The structure of the paper is as follows: Section 2 introduces necessary preliminaries of the paper and states the problem. Section 3 describes the contribution of the paper. An example is given in Section 4 and some conclusions are given in Section 5.

Notation: $\mathbb{R}^{m\times n}$ will denote the real matrices of size $m \times n$ and given a set of s symbolic variables $\xi = \{\xi_1, \ldots, \xi_s\}$, notation $\left[\binom{\xi}{q}\right]$ will denote the vector of all degree q monomials in variables ξ, comprised of $\frac{(s+q-1)!}{q!(s-1)!}$ elements, in a suitably prefixed ordering such as, for instance, lexicographical.

2. PRELIMINARIES AND PROBLEM STATEMENT

2.1 Nonlinear sector Takagi-Sugeno models

Consider a nonlinear discrete-time system

$$x_{t+1} = f(x_t, u_t), \tag{1}$$

with $x_t \in \mathbb{X} \subset \mathbb{R}^{n_x}$ and $u_t \in \mathbb{U} \subset \mathbb{R}^{n_u}$ being the model validity and input constraint region, being n_x and n_u the number of states and inputs, respectively. The nonlinear system in (1) can be **exactly** modelled using sector nonlinearity based on Takagi-Sugeno fuzzy models (Tanaka and Wang, 2001):

$$x_{t+1} = \sum_{i=1}^{\rho} \mu_i(x_t)(A_i x_t + B_i u_t) \tag{2}$$

being $\mu(x_t) = \{\mu_1(x_t), \ldots, \mu_\rho(x_t)\}$ a set of $\rho = 2^p$ nonlinear membership functions, with p the number of nonlinearities, where

$$\sum_{i=1}^{\rho} \mu_i(x_t) = 1, \quad 0 \le \mu_i(x_t) \le 1 \tag{3}$$

It is assumed, in (2), that nonlinearities of the model $\mu_i(x_t)$ are known functions and therefore they can be evaluated for a given state.

Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.07.091

2.2 Q-learning and optimal control

For a given arbitrary (stabilising) state-feedback policy, also known as "control law", $u_t := \pi(x_t)$ and initial state x_0, a scalar *value* function:

$$V^\pi(x_0) := \sum_{t=0}^{\infty} \gamma^t r(x_t, u_t) = \sum_{t=0}^{\infty} \gamma^t \left(x_t^T H_x x_t + u_t^T H_u u_t\right) \tag{4}$$

can be computed, being $\gamma \geq 1$ a decay rate, $r_t \equiv r(x_t, u_t)$ the so-called immediate reward or "cost" and $H_x \in \mathbb{R}^{n_x \times n_x}$ and $H_u \in \mathbb{R}^{n_u \times n_u}$ weighting matrices of the quadratic cost function.

The control objective is to design a controller $u_t = \pi(x_t)$ such that $V^\pi(x_t)$ is minimised. This objective is a generalization of the standard LQR/predictive to a nonlinear case. The parameter γ penalises costs later in time, so in a linear case it enforces a LQR with prescribed decay (a guaranteed discrete-time pole z faster than those with $|\bar{z}| = \gamma^{-1}$. Using the well-known Bellman equation, the *value* function can be computed by solving:
$V^\pi(x_t) = r_t + \gamma V^\pi(x_{t+1}) = r(x_t, \pi(x_t)) + \gamma V^\pi(f(x_t, \pi(x_t)))$, which can be seen as the expected return (cost) for a given state x_t if policy $\pi(x_t)$ were used at all future times. In addition to this, the *action-value* function $Q^\pi(x_t, u_t)$ is defined as the return for a given state and action, also known as Q-function:

$$Q^\pi(x_t, u_t) := r(x_t, u_t) + \gamma V^\pi(x_{t+1}). \tag{5}$$

For a given fixed policy $\pi(x_t)$, the following equivalence holds $V^\pi(x_t) = Q^\pi(x_t, \pi(x_t))$, and therefore the Q-function can be also expressed using the Bellman's equation:

$$Q^\pi(x_t, u_t) = r(x_t, u_t) + \gamma Q^\pi(x_{t+1}, \pi(x_{t+1})). \tag{6}$$

The optimal *value* function $V^{\pi^*}(x_t)$ and optimal control policy can be derived from the optimal *action-value* function $Q^{\pi^*}(x_t, u_t)$ using Bellman's principle of optimality:

$$V^{\pi^*}(x_t) := \min_{u_t} Q^{\pi^*}(x_t, u_t) \tag{7}$$

$$u_t^* := \pi^*(x_t) = \arg\min_{u_t} Q^{\pi^*}(x_t, u_t). \tag{8}$$

There are quite a few iterative algorithms to estimate such optimal policy in literature, based on dynamic programming (Bellman, 1957) and reinforcement learning (Sutton and Barto, 1998), such as value iteration, policy iteration, actor-critic setups, etc. Most of them reduce to Riccati equations (or iterations converging to the Riccati solution) for the linear case in model (1), see (Lewis et al., 2012).

Policy improvement. The Q-function is a key step in the so-called policy improvement step in the above algorithms. Indeed, it can be proved that, given a suboptimal policy $\pi(x_t)$ and its action-value function [1] $Q^\pi(x_t, u_t)$ and a policy improvement given by

$$\hat{u}_t := \hat{\pi}(x_t) := \arg\min_{u_t} Q^\pi(x_t, u_t), \tag{9}$$

we have:
$Q^\pi(x_t, \hat{\pi}(x_t)) = r(x_t, \hat{\pi}(x_t)) + \gamma Q^\pi(x_{t+1}, \pi(x_{t+1})) \leq Q^\pi(x_t, u_t)$. So, we have found an "improved" policy. Well-known argumentations allow to assert, too, that $V^{\hat{\pi}}(x_t) := Q^{\hat{\pi}}(x_t, \hat{\pi}(x_t)) \leq Q^\pi(x_t, \hat{\pi}(x_t)) \leq V^\pi(x_t)$ so such policy $\hat{\pi}(x_t)$ is preferable to the original one $\pi(x_t)$ actually at *every* instant. Also, optimal policies are a fixed point of Bellman if $\hat{\pi}(x_t) = \pi(x_t)$, then $\hat{\pi}(x_t) = \pi^*(x_t)$.

Linear discrete-time case. Let $f(x_t, u_t)$ be:

$$x_{t+1} = Ax_t + Bu_t,$$

under a state feedback $u_t = \pi(x_t) = -K^\pi x_t$. It can be proved that the Q-function is quadratic:

$$Q^\pi(x_t, u_t) := \begin{bmatrix} x_t \\ u_t \end{bmatrix}^T \begin{bmatrix} S_{xx}^\pi & S_{xu}^\pi \\ S_{ux}^\pi & S_{uu}^\pi \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix} = \tag{10}$$

$$= \begin{bmatrix} x_t \\ u_t \end{bmatrix}^T \begin{bmatrix} H_x + \gamma A^T P^\pi A & \gamma A^T P^\pi B \\ \gamma B^T P^\pi A & \gamma B^T P^\pi B + H_u \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix},$$

where P^π is obtained from the Lyapunov equation associated to the feedback gain:

$$\gamma (A - BK^\pi)^T P^\pi (A - BK^\pi) - P^\pi + H_x + (K^\pi)^T H_u K^\pi = 0.$$

By applying $\frac{\partial Q^\pi(x_t, u_t)}{\partial u_t} = 0$ to (10) we get $\hat{u}_t$ in (9):

$$\hat{u}_t = -\left(S_{uu}^\pi\right)^{-1} S_{ux}^\pi x_t$$

as a basis for the policy improvement. If P^π were the optimal solution (Riccati equation with prescribed decay) then $K^\pi = \left(S_{uu}^\pi\right)^{-1} S_{ux}^\pi$ would be equal to the optimal LQR gain.

Nonlinear case. In the nonlinear case, no explicit solution for $V^{\pi^*}(x_t)$ and u_t^* can be found in general in (7) and (8), only upper bounds (see later in this section), with LMI techniques, unrelated to learning, and only in specific cases. Hence, most of Q-learning approaches propose to learn a specific parametrization of $Q^\pi(x_t, u_t)$ using regressors (Lewis et al., 2012), i.e.: using a linear parametrization:

$$Q^\pi(x_t, u_t) \approx (\omega^\pi)^T \varphi(x_t, u_t) \tag{11}$$

with $\varphi(x_t, u_t) : \mathbb{X} \times \mathbb{U} \mapsto \mathbb{R}^m$ being a known set of m regression functions and $\omega^\pi \in \mathbb{R}^{m \times 1}$ the weights to be learned. The linear case is that in which $\varphi(x_t, u_t)$ is the vector with all the second order monomials of the combined input-state vector:

$$\varphi(x_t, u_t) = \left[\begin{pmatrix} \{x_t, u_t\} \\ 2 \end{pmatrix} \right]$$

Policy evaluation with Temporal Difference. Given a policy $\pi(x_t)$ and a Q-function based on (11), and a series of N data points collected from an experiment [2]

$$\mathcal{D} = \{\{x_1, u_1\}, \{x_2, u_2\}, \ldots, \{x_N, u_N\}\}$$

the weights in $Q^\pi(x_t, u_t)$ can be learned from data using Temporal-Difference (TD) methods (Lewis et al., 2012; Lewis and Vrabie, 2009; Busoniu et al., 2010) using equation (6):

$$(\omega^\pi)^T (\varphi(x_t, u_t) - \gamma \varphi(\bar{x}_{t+1}, \pi(\bar{x}_{t+1})) = r(x_t, u_t)$$

where $\bar{x}_{t+1} = f(x_t, u_t)$. Indeed, $R \in \mathbb{R}^{N \times 1}$ is the inmediate rewards (computed from each data point) vector, and matrices $\Phi \in \mathbb{R}^{N \times m}$ and $\Phi^{\pi+} \in \mathbb{R}^{N \times m}$ whose rows

[1] Actually in most cases, an approximation of it.

[2] Ideally $\mathcal{D}$ should cover the whole space $\mathbb{X}$ and $\mathbb{U}$ with a random combined state-input.