Automatica 64 (2016) 94–104

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Model-based reinforcement learning for approximate optimal regulation☆

Rushikesh Kamalapurkar, Patrick Walters, Warren E. Dixon

Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, USA

ARTICLE INFO

Article history:
Received 10 May 2013
Received in revised form
23 July 2015
Accepted 2 October 2015
Available online 7 December 2015

Keywords:
Model-based reinforcement learning
Concurrent learning
Simulated experience
Data-based control
Adaptive control
System identification

ABSTRACT

Reinforcement learning (RL)-based online approximate optimal control methods applied to deterministic systems typically require a restrictive persistence of excitation (PE) condition for convergence. This paper develops a concurrent learning (CL)-based implementation of model-based RL to solve approximate optimal regulation problems online under a PE-like rank condition. The development is based on the observation that, given a model of the system, RL can be implemented by evaluating the Bellman error at any number of desired points in the state space. In this result, a parametric system model is considered, and a CL-based parameter identifier is developed to compensate for uncertainty in the parameters. Uniformly ultimately bounded regulation of the system states to a neighborhood of the origin, and convergence of the developed policy to a neighborhood of the optimal policy are established using a Lyapunov-based analysis, and simulation results are presented to demonstrate the performance of the developed controller.

1. Introduction

Reinforcement learning (RL) enables a cognitive agent to learn desirable behavior from interactions with its environment. In control theory, the desirable behavior is typically quantified using a cost function, and the control problem is formulated as the desire to find the optimal policy that minimizes a cumulative cost. RL techniques for discrete time systems are inherently model-free, and hence, have been a prime focus of research over the past few decades (Kaelbling, Littman, & Moore, 1996).

Recently, various RL-based techniques have been developed to approximately solve optimal control problems for continuous-time and discrete-time deterministic systems (Al-Tamimi, Lewis, & Abu-Khalaf, 2008; Bhasin et al., 2013; Dierks, Thumati, & Jagannathan, 2009; Doya, 2000; Lewis & Vrabie, 2009; Mehta & Meyn, 2009; Padhi, Unnikrishnan, Wang, & Balakrishnan, 2006; Vamvoudakis & Lewis, 2010; Zhang, Cui, & Luo, 2013; Zhang, Cui, Zhang, & Luo, 2011; Zhang, Liu, Luo, & Wang, 2013). The approximate solution is facilitated via value function approximation, where the optimal policy is computed based on an estimate of the value function.

Methods that seek online solutions to optimal control problems are comparable to adaptive control (cf., Bhasin et al., 2013; He & Jagannathan, 2007; Padhi et al., 2006; Vamvoudakis & Lewis, 2010; Zhang et al., 2013; Zhang, Wei, & Luo, 2008 and the references therein). In adaptive control, the estimates for the uncertain parameters in the plant model are updated using the tracking error as a performance metric; whereas, in online RL-based techniques, estimates for the uncertain parameters in the value function are updated using the Bellman error (BE) as a performance metric. Typically, to establish regulation or tracking, adaptive control methods do not require the adaptive estimates to convergence to the true values. However, convergence of the RL-based controller to a neighborhood of the optimal controller requires convergence of the parameter estimates to a neighborhood of their ideal values.

Parameter convergence has been a focus of research in adaptive control for several decades. It is common knowledge that least squares and gradient descent-based update laws generally require persistence of excitation (PE) in the system state for convergence of the parameter estimates. Modification schemes such as projection algorithms, σ-modification, and e-modification are used to guarantee boundedness of parameter estimates and overall system stability; however, these modifications do not guarantee parameter convergence unless the PE condition is satisfied (Ioannou & Sun, 1996; Narendra & Annaswamy, 1987, 1989; Sastry & Bodson, 1989).

☆ The material in this paper was partially presented at the 52nd IEEE Conference on Decision and Control, December 10–13, 2013, Florence, Italy. This paper was recommended for publication in revised form by Associate Editor Antonio Loria under the direction of Editor Andrew R. Teel.

E-mail addresses: rkamalapurkar@ufl.edu (R. Kamalapurkar), walters8@ufl.edu (P. Walters), wdixon@ufl.edu (W.E. Dixon).

http://dx.doi.org/10.1016/j.automatica.2015.10.039
0005-1098/© 2015 Elsevier Ltd. All rights reserved.

In RL-based approximate online optimal control, the Hamilton–Jacobi–Bellman (HJB) equation along with an estimate of the state derivative (cf. Bhasin et al., 2013; Mehta & Meyn, 2009), or an integral form of the HJB equation (cf. Vrabie, 2010) is utilized to approximately evaluate the BE along the system trajectory. The BE, evaluated at a point, provides an indirect measure of the quality of the estimate of the value function evaluated at that point. Hence, the unknown value function parameters are updated based on evaluation of the BE along the system trajectory. Such weight update strategies create two challenges for analyzing convergence. The system states need to satisfy PE, and the system trajectory needs to visit enough points in the state space to generate a good approximation of the value function over the entire domain of operation. These challenges are typically addressed in the related literature (cf. Al-Tamimi, Lewis, & Abu-Khalaf, 2007; Bhasin et al., 2013; Kiumarsi, Lewis, Modares, Karimpour, & Naghibi-Sistani, 2014; Lewis & Vrabie, 2009; Modares & Lewis, 2014; Modares, Lewis, & Naghibi-Sistani, 2013, 2014; Vamvoudakis & Lewis, 2010, 2011; Vamvoudakis, Lewis, & Hudas, 2012) by adding an exploration signal to the control input to ensure sufficient exploration of the domain of operation. However, no analytical methods exist to compute the appropriate exploration signal when the system dynamics are nonlinear.

The aforementioned challenges arise from the restriction that the BE can only be evaluated along the system trajectories. In particular, the integral BE is meaningful as a measure of quality of the value function estimate only if it is evaluated along the system trajectories, and state derivative estimators can only generate numerical estimates of the state derivative along the system trajectories. Recently, Modares et al. (2014) demonstrated that experience replay can be used to improve data efficiency in online approximate optimal control by reuse of recorded data. However, since the data needs to be recorded along the system trajectory, the system trajectory under the designed approximate optimal controller needs to provide enough excitation for learning. In general, such excitation is not available; hence, the simulation results in Modares et al. (2014) are generated using an added probing signal.

In this paper, and in our preliminary work in Kamalapurkar, Walters, and Dixon (2013), a different approach is used to improve data efficiency by observing that if the system dynamics are known, the state derivative, and hence, the BE can be evaluated at any desired point in the state space. Unknown parameters in the value function can therefore be adjusted based on least square minimization of the BE evaluated at any number of arbitrary points in the state space. For example, in an infinite horizon regulation problem, the BE can be computed at points uniformly distributed in a neighborhood around the origin of the state space. The results of this paper indicate that convergence of the unknown parameters in the value function is guaranteed provided the selected points satisfy a rank condition. Since the BE can be evaluated at any desired point in the state space, sufficient exploration can be achieved by appropriately selecting the points to cover the domain of operation. If the system dynamics are partially unknown, an approximation to the BE can be evaluated at any desired point in the state space based on an estimate of the system dynamics.

If each new evaluation of the BE along the system trajectory is interpreted as gaining experience via exploration, the use of a model to evaluate the BE at an unexplored point in the state space can be interpreted as a simulation of the experience. Learning based on simulation of experience has been investigated in results such as Abbeel, Quigley, and Ng (2006); Atkeson and Schaal (1997); Deisenroth (2010); Deisenroth and Rasmussen (2011); Mitrovic, Klanke, and Vijayakumar (2010); Singh (1992) for stochastic model-based RL; however, these results solve the optimal control problem off-line in the sense that repeated learning trials need to be performed before the algorithm learns the controller, and system stability during the learning phase is not analyzed. This paper furthers the state of the art for nonlinear, control affine plants with linearly parameterizable (LP) uncertainties in the drift dynamics by providing an online solution to deterministic infinite horizon optimal regulation problems. In this paper, a CL-based parameter estimator is developed to exponentially identify the unknown parameters in the system model, and the parameter estimates are used to implement simulation of experience by extrapolating the BE.

The main contributions of this paper include a novel implementation of model-based RL in deterministic nonlinear systems and a detailed stability analysis that establishes simultaneous online identification of system dynamics and online approximate learning of the optimal controller, while maintaining system stability. Simulation results are provided that demonstrate the approximate solution of infinite horizon optimal regulation problems online for inherently unstable nonlinear systems with uncertain drift dynamics. The simulations also demonstrate that the developed method can be used to implement RL without the addition of a probing signal.

2. Problem formulation

Consider a control affine nonlinear dynamical system[1]

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

where $x : \mathbb{R}_{\geq t_0} \to \mathbb{R}^n$ denotes the system state trajectory, $u : \mathbb{R}_{\geq t_0} \to \mathbb{R}^m$ denotes the control input, $f : \mathbb{R}^n \to \mathbb{R}^n$ denotes the drift dynamics, and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ denotes the control effectiveness. In the following, the notation $\phi^u(t; t_0, x^o)$ denotes a trajectory of the system in (1) under the controller u with the initial condition $x^o \in \mathbb{R}^n$ and initial time $t_0 \in \mathbb{R}_{\geq 0}$.[2] The objective is to solve the infinite horizon optimal regulation problem online, i.e., to find the optimal policy $u^* : \mathbb{R}^n \to \mathbb{R}^m$ defined as

$$u^*\left(x^o\right) \triangleq \underset{u(\tau)\in U|\tau\in\mathbb{R}_{\geq t}}{\arg\min} \int_t^\infty r\left(\phi^u\left(\tau; t, x^o\right), u(\tau)\right) d\tau, \tag{2}$$

while regulating the system states to the origin.[3] In (2), $U \in \mathbb{R}^m$ denotes the action space and $r : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ denotes the instantaneous cost defined as $r(x^o, u^o) \triangleq x^{oT}Qx + u^{oT}Ru^o$, where $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are constant positive definite symmetric matrices. The class of nonlinear systems considered in this paper is characterized by the following assumption.

Assumption 1. The drift dynamics f is an unknown, LP locally Lipschitz function such that $f(0) = 0$, and the control effectiveness g is a known bounded locally Lipschitz function. Furthermore, $f' : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is continuous, where $(\cdot)'$ denotes the partial derivative with respect to the first argument.

A closed-form solution to the optimal control problem is formulated in terms of the optimal value function $V^* : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ defined as

$$V^*\left(x^o\right) \triangleq \min_{u(\tau)\in U|\tau\in\mathbb{R}_{\geq t}} \int_t^\infty r\left(\phi^u\left(\tau; t, x^o\right), u(\tau)\right) d\tau. \tag{3}$$

[1] For notational brevity, unless otherwise specified, the domain of all the functions is assumed to be $\mathbb{R}_{\geq 0}$, where $\mathbb{R}_{\geq a}$ denotes the interval $[a, \infty)$. The notation $\|\cdot\|$ denotes the Euclidean norm for vectors and the Frobenius norm for matrices. The notation $(\cdot)^o$ denotes arbitrary variables.

[2] Whenever the initial time and state are implied or unimportant, a trajectory of the system in (1) evaluated at time t will be denoted by $x(t)$.

[3] The definition in (2) implicitly assumes existence of the optimal policy.