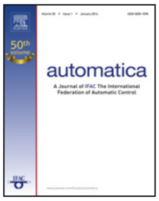
Automatica 50 (2014) 475–489

Contents lists available at ScienceDirect

Automatica

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



On integral generalized policy iteration for continuous-time linear quadratic regulations[☆]

Jae Young Lee[a], Jin Bae Park[a,1], Yoon Ho Choi[b]

[a] Department of Electrical and Electronic Engineering, Yonsei University, 262 Seongsanno, Seodaemun-gu, Seoul, Republic of Korea

[b] Department of Electronic Engineering, Kyonggi University, 94-6 Yiui-dong, Yeongtong-gu, Suwon, Kyonggi-Do, Republic of Korea

ARTICLE INFO

Article history:
Received 27 September 2012
Received in revised form
3 May 2013
Accepted 20 November 2013
Available online 30 December 2013

Keywords:
LQR
Generalized policy iteration
Reinforcement learning
Adaptive control
Optimization under uncertainties

ABSTRACT

This paper mathematically analyzes the integral generalized policy iteration (I-GPI) algorithms applied to a class of continuous-time linear quadratic regulation (LQR) problems with the unknown system matrix A. GPI is the general idea of interacting policy evaluation and policy improvement steps of policy iteration (PI), for computing the optimal policy. We first introduce the update horizon $\hbar$, and then show that (i) all of the I-GPI methods with the same $\hbar$ can be considered equivalent and that (ii) the value function approximated in the policy evaluation step monotonically converges to the exact one as $\hbar \to \infty$. This reveals the relation between the computational complexity and the update (or time) horizon of I-GPI as well as between I-PI and I-GPI in the limit $\hbar \to \infty$. We also provide and discuss two modes of convergence of I-GPI; I-GPI behaves like PI in one mode, and in the other mode, it performs like value iteration for discrete-time LQR and infinitesimal GPI ($\hbar \to 0$). From these results, a new classification of the integral reinforcement learning is formed with respect to $\hbar$. Two matrix inequality conditions for stability, the region of local monotone convergence, and data-driven (adaptive) implementation methods are also provided with detailed discussion. Numerical simulations are carried out for verification and further investigations.

1. Introduction

In the field of computational intelligence, generalized policy iteration (GPI) is the general idea of interacting the two consecutive steps of (iterative) policy iteration (PI) or actor-critic methods, for computing the optimal policy in a Markov decision process (MDP). The respective two revolving steps are *policy evaluation*, making the value function in critic consistent with the current policy, and *policy improvement*, making the policy in actor greedy with respect to the current value function (Sutton & Barto, 1998). This general idea allows one of these two steps to be performed without completing the other step *a priori*. Almost all reinforcement learning (RL) and approximate dynamic programming (DP) methods are well described by this idea of GPI including actor-critic methods and modified PI (Bertsekas & Tsitsiklis, 1996; Puterman & Shin, 1978; Sutton & Barto, 1998).

Modified PI, classified as a class of GPI methods (Sutton & Barto, 1998), was first formulated by Puterman and Shin (1978) and van Nunen (1976) in finite MDP frameworks. It was created by approximating the policy evaluation of the exact PI by the finite k-number of Bellman fixed point iterations; the exact PI ($k \to \infty$) and value iteration (VI) ($k = 1$) fall into special cases of this (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998). Here, the natural number k, called the *iteration horizon* of GPI in this paper, mediate a trade-off between the computational complexity (large k) and the approximation error (small k). For all $k \in \mathbb{N} \cup \{\infty\}$, the convergence to the optimal solution was proved with the connection to the DP operator and its properties (Bertsekas & Tsitsiklis, 1996).

Based on the results of finite MDP frameworks, extensive research has been carried out to develop the RL and approximate DP algorithms for continuous-state dynamical systems (CSDS) in both discrete-time (DT) domain (Al-Tamimi, 2007; Jiang & Jiang, 2010; Lendelius, 1997; Prokhorov & Wunsch, 1997; Si, Barto, Powell, & Wunsch, 2004; Wang, Liu, Wei, Zhao, & Jin, 2012; Webos, 1992; Zhang, Huang, & Lewis, 2009) and recently, continuous-time (CT) domain (Bhasin et al., 2013; Doya, 2000; Hanselmann, Noakes, & Zaknich, 2007; Lee, Park, & Choi, 2010, 2012; Vamvoudakis & Lewis, 2010; Vrabie, 2009; Vrabie & Lewis, 2009). Lewis and Vrabie (2009) and Wang, Zhang, and Liu (2009) performed recent

[☆] This work has been supported by Institute of BioMed-IT, Energy-IT and Smart-IT Technology (BEST), a Brain Korea 21 plus program, Yonsei University. The material in this paper was partially presented at the 50th IEEE Conference on Decision and Control (CDC) and European Control Conference, December 12–15, 2011, Orlando, Florida, USA and the 2013 American Control Conference, June 17–19, 2013, Washington, DC, USA. This paper was recommended for publication in revised form by Associate Editor Shuzhi Sam Ge under the direction of Editor Miroslav Krstic.

E-mail addresses: jyounglee@yonsei.ac.kr (J.Y. Lee), jbpark@yonsei.ac.kr (J.B. Park), yhchoi@kyonggi.ac.kr (Y.H. Choi).

[1] Tel.: +82 2 2123 2773; fax: +82 2 362 4539.

0005-1098/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.automatica.2013.12.009

surveys about these algorithms. In these cases, however, most of the research was focused only on the two extreme cases, namely, PI ($k \to \infty$, *maximum computational complexity*) and VI ($k = 1$, *maximum approximation error*). In those studies, the development of VI for CSDS was parallel to that for a finite MDP (Al-Tamimi, 2007; Lee et al., 2010; Lewis & Vrabie, 2009; Prokhorov & Wunsch, 1997; Si et al., 2004; Vrabie, 2009; Wang et al., 2009; Webos, 1992), but PI for CSDS additionally needs the assumption of an initial stabilizing policy to guarantee its stability and convergence (Lee et al., 2012; Lewis & Vrabie, 2009; Vrabie, 2009; Wang et al., 2009). Moreover, there are two different ways of implementing the policy evaluation of PI (Lewis & Vrabie, 2009; Vrabie, 2009)—one is based on the Bellman's fixed point iterations similar to the finite MDP case, resulting in high computational complexity due to the extremely large k (theoretically, $k \to \infty$), and the other uses the difference regression vectors which are less likely excited than those of VI and thereby decrease the computability and accuracy of the value function. Therefore, compared with the VI methods, the PI algorithms for CSDS are computationally expensive, regardless of which implementation method is used.

For CSDS, the process of solving a given optimal control problem generally falls into that of computing the solution of the underlying Hamilton–Jacobi–Bellman (HJB) equation whose analytical solution is difficult to obtain in general. In the case of the PI and VI, the HJB equation is iteratively solved by revolving policy evaluation and improvement steps, performed by critic and actor networks, respectively. In this process, the Lyapunov function associated with the current policy is evaluated or approximated by critics in the (approximate) policy evaluation step, and the policy is updated by actor in the policy improvement step, based on the current (approximated) Lyapunov function (Al-Tamimi, 2007; Lewis & Vrabie, 2009; Si et al., 2004; Vrabie, 2009; Wang et al., 2009). While PI finds the exact Lyapunov function by policy evaluation, VI approximates the Lyapunov function by only one step recursion.

For linear systems, the HJB equation becomes the well-known algebraic Riccati equation (ARE), and the above two steps of PI and VI can be considered as the process of solving the associated Lyapunov matrix equation/recursion and updating the policy by using the matrix solution (Al-Tamimi, 2007; Jiang & Jiang, 2010; Lendelius, 1997; Lee et al., 2010, 2012; Lewis & Vrabie, 2009; Vrabie, 2009; Zhang et al., 2009). In fact, this kind of iterative method was already developed independently, with a number of analyses on convergence, stability, and computational complexity in the fields of control engineering and numerical analysis (Feitzinger, Hylla, & Sachs, 2009; Hewer, 1971; Kleinman, 1968; Lancaster & Rodman, 1995; Stoorvogel & Weeren, 1994). From these results, a number of control and learning schemes based on PI or VI were also analyzed by showing the equivalence of each to one of the existing iterative methods. For PI methods, which exactly evaluate the Lyapunov matrix solution, it was shown that in the case of linear quadratic regulations (LQR), they are equivalent to Newton methods and thereby guarantee the stability and 2nd-order monotone decreasing convergence (Jiang & Jiang, 2010; Lee et al., 2012; Lewis & Vrabie, 2009; Vrabie, 2009). In the case of DT VI, the equivalence to the Lyapunov matrix recursions also provides convergence to the optimal solution (Al-Tamimi, 2007; Lendelius, 1997; Lewis & Vrabie, 2009; Zhang et al., 2009); the convergence is monotone and increasing for LQR case. Similar analytical results also exist for nonlinear PI and VI algorithms (Al-Tamimi, 2007; Lewis & Vrabie, 2009; Vrabie, 2009).

The concept of GPI in DT CSDS was introduced by Lewis and Vrabie (2009) from the perspectives of modified PI. Similar to GPI in MDP frameworks, VI ($k = 1$) and PI ($k \to \infty$) for DT CSDS are two extreme cases of this GPI. On the other hand, a number of actor-critic methods for input-affine CSDS have been proposed in CT domain from the GPI viewpoint—concurrent actor-critic learning (Bhasin et al., 2013; Hanselmann et al., 2007; Vamvoudakis & Lewis, 2010) and modified PI (Vrabie, 2009; Vrabie & Lewis, 2009). The GPI method we have focused on in this paper is the modified PI given by Vrabie and Lewis (2009). This GPI method, together with the related PI and VI as two special cases, belongs to a class of algorithms known as integral (or interval) RL (I-RL). These I-RL algorithms iteratively perform (approximate) policy evaluation and improvement steps *without knowing the system drift dynamics*, using the integral reinforcement signal made by observing the cost during the *finite time horizon* T_s (Lewis & Vrabie, 2009; Vrabie, 2009). On the contrary, the concurrent actor-critic methods require either full-knowledge about the system dynamics (Hanselmann et al., 2007; Vamvoudakis & Lewis, 2010) or an associated system identifier (Bhasin et al., 2013). In this paper, the I-RL algorithms based on GPI, PI, and VI methods for CT CDSD will be called *integral GPI* (I-GPI), *integral PI* (I-PI), and *integral VI* (I-VI), respectively.

Among the I-RL methods, considerable efforts have been made on the analysis of I-PI in terms of stability, monotonicity, and convergence. The analyses of I-PI are based on the equivalence to certain numerical iteration methods. As mentioned above, it was proved that in the case of LQR, I-PI is equivalent to Kleinman (1968)'s Newton method which monotonically improves the policy by iterations and guarantees the global stability and 2nd-order convergence (Vrabie, 2009). Further analysis and extensions can be found in Lee et al. (2012). In the case of I-VI for LQR, the stability and convergence conditions were investigated based on matrix operators (Lee et al., 2010; Vrabie, 2009). For the policy evaluation step of I-GPI, Vrabie and Lewis (2009) proved that, under an admissible policy, the value function approximated by the k-number of Bellman's fixed-point iterations converges to the exact one as $k \to \infty$. The proof was based on the DP operator and its properties, similar to the modified PI in finite MDP frameworks. To the best of the authors' knowledge, however, there is no further analysis of the I-GPI algorithms, even for the LQR case in terms of stability, monotone convergence, and equivalences.

In this paper, we mathematically analyze I-GPIs applied to *CT LQR problems* with unknown system matrix A. While the I-GPI method given by Vrabie and Lewis (2009) assumes an initial stabilizing policy, ours does not for analytical purposes. The update horizon $\hbar$, first introduced in this paper as the product of the iteration and time horizons ($\hbar := kT_s$), plays a central role in the analysis. The main contributions of this paper can be summarized as follows:

1. From the process of re-derivations of I-GPI, we show that the I-GPI algorithms that use the same $\hbar$ are all equivalent in the iteration domain. This shows that for the same $\hbar$, the computational complexity due to large k can be lessened by increasing the time horizon T_s.
2. For policy evaluation recursion of I-GPI, a sub-iteration in each policy evaluation step, we provide monotone convergence properties with respect to the update horizon $\hbar$, which imply the equivalence of I-PI and the I-GPI methods in the limit $\hbar \to \infty$ under an initial stabilizing policy. These are the extensions of the work of Vrabie and Lewis (2009), where only the convergence in the limit of the iteration horizon ($k \to \infty$) was investigated.
3. A number of (matrix) inequality conditions are provided for *closed-loop stability* and/or *global/local monotone convergence* of I-GPI methods. Here, two modes of global convergence are considered—one, called PI-mode of convergence, behaves like PI, and the other, called VI-mode of convergence, occurs for sufficiently small $\hbar$ and acts like VI for DT LQR and infinitesimal GPI ($\hbar \to 0$). Based on these two modes of convergence and the properties of I-GPI regarding the update horizon $\hbar$, a new spectral classification of I-RL algorithms is established with respect