2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

On Generalized Policy Iteration for Continuous-Time Linear Systems

Jae Young Lee, Tae Yoon Chun, Jin Bae Park*, and Yoon Ho Choi

Abstract— **This paper investigate the mathematical properties of generalized policy iteration (GPI) applied to a class of continuous-time linear systems with unknown internal dynamics. GPI is a class of dynamic programming (DP) method to solve an optimal control problem by using two consecutive steps—policy evaluation and policy improvement. We first provide several formula equivalent to GPI, and as a result, reveal its relations to linear quadratic optimal control problems and the fact that the computational complexity due to backup operations in policy evaluation steps can be lessened by increasing the time horizon of GPI. A variety of local stability and convergence criteria is also provided with the connection to the convergence speed. Finally, several numerical simulations are performed to verify the results.**

I. Introduction

In the field of computational intelligence, generalized policy iteration (GPI), together with policy iteration (PI) and value iteration (VI), are well-known dynamic programming (DP) algorithms, with extensive practical applications [1]–[3], for computing optimal policies for a finite Markov decision process (MDP) iteratively [1]. These algorithms are closely related to reinforcement learning (RL) [1], and consist of two consecutive interactive steps, the one called *policy evaluation* making the value function approximately consistent with the current policy, and the other called *policy improvement* making the policy greedy in terms of the consistent value function [1]–[3].

The key difference among these algorithms lies in the policy evaluation step—PI evaluates the exact value function with respect to the current policy, which requires the infinite number of iterations; VI executes only one-step recursion in policy evaluation, which decreases the computational burden commonly arising in PI, but introduces approximation error of the evaluated value function in return [1]–[3]. Meanwhile, GPI lies between PI and VI—it takes a number k of iterations ($1 \leq k \leq \infty$) in policy evaluation to evaluate an approximate value function, making a tradeoff between the accuracy and computational complexity depending on how large k is. Here, we refer to k as the *iteration horizon* of GPI. Depending this iteration horizon k, PI ($k = \infty$) and VI ($k = 1$) can be considered the special case of GPI algorithm.

Based on the results in MDP framework, extensive researches have been carried out on extending those DP algorithms to the dynamic systems in discrete-time (DT) domain [3]–[6] at first, and later, in continuous-time (CT) framework [7]–[13] (see [4] and [6] for a survey). Among these extensions to DT and CT dynamic systems [3]–[13] (to the best authors' knowledge), there is only one GPI technique, given by Vrabie *et al.* [11], which belongs to a class of algorithms named as interval (or integral) RL. These I-RL iteratively performs policy evaluation and improvement steps by observing the cost during the *finite time horizon* T, to solve a class of optimal control problems regarding CT dynamic systems with unknown internal dynamics [6], [11]. According to the spirit of the algorithms in a finite MDP above, these I-RL methods can be classified into PI [6], [10], VI [6], [9], and GPI [11], where we call these algorithms in this paper *integral PI* (I-PI), *integral VI* (I-VI), and *integral GPI* (I-GPI), respectively.

In addition to the applicability to CT dynamic systems with unknown internal dynamics, the advantage of these I-RLs over the others is that the stability and convergence properties are well-analyzed [6], [10], [11], [13]. In case of linear quadratic regulation (LQR), it was proven in [10] that I-PI is equivalent to the Kleinman's Newton formula which guarantees the global stability and convergence to the optimal solution, with local 2^{nd}-order convergence [14]. In case of I-VI for LQR, the local stability and convergence conditions were investigated by Lee *et al.* [13], with a generalized framework. For I-GPI, it was proven in [11] that under the admissible policy assumption, the value function approximated by k-number of iterations in the policy evaluation step converges to the exact one as $k \to \infty$. However, to the best authors' knowledge, the stability, monotonicity, and convergence of I-GPI as well as its relations to the target optimal control problems were not explored even for LQR case. Moreover, over the all I-RL algorithms, there exists no analysis about how the iteration horizon k and time horizon T affect the accuracy and computational complexity of the policy evaluations.

This paper deeply focuses on the I-GPI algorithms applied to *LQR problems* and provides various mathematical results. First, the relationships between the iteration and time horizon k and T are investigated with the connection to the convergence, accuracy, and computational complexity of the policy evaluation steps of I-GPI. Second, various local stability, monotonicity, and convergence criteria are suggested for I-GPI; Third, several useful equivalent formula of the I-GPI are provided with their strong connections to LQR. Finally, numerical simulations are performed to verify these results.

This research was supported by the Brain Korea 21 Project in 2011, and by the Ministry of Knowledge Economy (MKE), Project # 2009T100100609, "Development for 22.9kV Power Cable Fault Diagnosis System and Traceability Management System Based on TFDR".

J. Y. Lee and J. B. Park are with Department of Electrical and Electronic Engineering, Yonsei University, Shinchon-Dong, Seodaemum-Gu, Seoul 120-749, Korea (E-mail: jyounglee@yonsei.ac.kr, jbpark@yonsei.ac.kr)

Y. H. Choi is with Department of Electronic Engineering, Kyonggi University, Suwon, Kyonggi-Do 443-760, Korea (E-mail: yhchoi@kyonggi.ac.kr)

* Corresponding author.

978-1-61284-799-3/11/$26.00 ©2011 IEEE

Notations: In the sequel, $\mathbb{M}^{m\times n}$ denotes the set of all $m\times n$ matrices; $\mathbb{M}_P^{n\times n}$ (resp. $\mathbb{M}_{PS}^{n\times n}$) is the set of all $n\times n$ positive definite (resp. semidefinite) matrices. For any matrix $M\in\mathbb{M}^{m\times n}$ and vector $x\in\mathbb{R}^n$, M' is the transpose of M; $\|M\|$ and $\|x\|$ denote the spectral norm (a maximum singular value) of M and the Euclidean norm $(x'x)^{1/2}$ of x, respectively.

II. Related Topics on LQR

In this section, we focus on the topics on LQR, which is closely related to the GPI algorithm—the one concerning the value function $V_u(x_t,t)$ with a stabilizing policy u, and the other concerning Bellman's optimality principle with the DP operator. In the first place, we state the following lemma, which will be extensively employed throughout the paper:

Lemma 1: For any matrices $X\in\mathbb{M}^{n\times n}$ and $Y\in\mathbb{M}^{n\times n}$, the following integral formula holds for all $T>0$:

$$e^{X'T}Ye^{XT}-Y=\int_0^T e^{X'\tau}(X'Y+YX)e^{X\tau}\,d\tau.$$

■

In addition, if X is assumed Hurwitz, then This can be simplified as

$$-Y=\int_0^\infty e^{X'\tau}(X'Y+YX)e^{X\tau}\,d\tau \tag{1}$$

by letting $T\to\infty$. Together with Lemma 1, this equation will be used to explain the connection of LQR and GPI algorithm.

A. Value Function with a Stabilizing Policy

Now, consider the following CT linear system ($t\geq 0$):

$$\dot{x}_t=Ax_t+Bu_t, \tag{2}$$

for a state $x_t\in\mathbb{R}^n$, a control input $u_t\in\mathbb{R}^m$, and the matrices $A\in\mathbb{M}^{n\times n}$ and $B\in\mathbb{M}^{n\times m}$, with the infinite-horizon quadratic value function

$$V_u(x_t,t)=\int_t^\infty x_\tau' S x_\tau+u_\tau' R u_\tau\,d\tau \tag{3}$$

where $S\in\mathbb{M}_{PS}^{n\times n}$ and $R\in\mathbb{M}_P^{m\times m}$. Here, throughout the paper, $u(t)$, u_t, and simply u will be used interchangeably for the input of the system (2) and the triple $(A,B,S^{1/2})$ is assumed to be stabilizable and detectable.

Let $u=-Kx$ be any policy for the system (2) and A_K its corresponding closed loop matrix $A-BK$. Defining Q_K for a policy K as $Q_K:=S+K'RK$ for simplicity, then, we can represent $V_u(x_t,t)$ in terms of Q_K as

$$V_u(x_t,t)=x_t'\left(\int_t^\infty e^{A_K'(\tau-t)}Q_K\,e^{A_K(\tau-t)}\,d\tau\right)x_t=x_t'P_Kx_t \tag{4}$$

where P_K is defined as $P_K:=\displaystyle\int_0^\infty e^{A_K'\tau}Q_K\,e^{A_K\tau}\,d\tau.$ (5)

If $u=-Kx$ is a stabilizing policy, then, the value function (3) is finite [14], making the problem feasible. Now, by applying (1) to P_K with $X=A_K$, (5) can be rewritten as

$$\int_0^\infty e^{A_K'\tau}Ric_K(P_K)e^{A_K\tau}\,d\tau=0 \tag{6}$$

for a stabilizing $u=-Kx$, where $Ric_K(P_K)$ is defined as $Ric_K(P_K):=A_K'P_K+P_KA_K+Q_K$. Note that (6) always holds for all stabilizing K and P_K. This implies the pair (K,P_K) always satisfies the Lyapunov equation $Ric_K(P_K)=0$. Such P_K always exists uniquely for any given stabilizing K and $Q_K\in M_P^{n\times n}$ [15]. Therefore, for any stabilizing K, one can always find the corresponding value function $V_u(x)=x^TP_Kx$ by solving the Lyapunov equation $Ric_K(P_K)=0$.

B. DP Operator & Optimality Principle

Regarding the system dynamics (2), we define the dynamic programming operator $\mathscr{T}_K:X\to X$ on the space X of the continuous functionals $V(x):\mathbb{R}^n\to\mathbb{R}$ at fixed time $t\geq 0$ as

$$\mathscr{T}_KV(x):=\int_t^{t+T}x_\tau'Q_Kx_\tau\,d\tau+V(x_{t+T}). \tag{7}$$

where the trajectories of x_t are generated by the system (2) with a given control $u=-Kx$. We also define $(\mathscr{T}_K)^2$ as $(\mathscr{T}_K)^2V(x):=\mathscr{T}_K[\mathscr{T}_KV(x)]$, and so does $(\mathscr{T}_K)^k$ at fixed time $t\geq 0$ for any $k\in\mathbb{N}$. This operator simplifies the mathematical statements related with the optimality principle and I-GPI algorithm. Moreover, it also possesses the following useful mathematical properties:

Lemma 2: Consider the system dynamics (2) with $u=-Kx$ and a continuous functional of the form $V(x)=x^TPx$. Then, we have

$$\mathscr{T}_KV(x)=x_t'\left(P+\int_0^T e^{A_K'\tau}Ric_K(P)e^{A_K\tau}\,d\tau\right)x_t, \tag{8}$$

$$\begin{aligned}\frac{d}{dt}\mathscr{T}_KV(x)=&x_t'\left(e^{A_K'T}Ric_K(P)e^{A_KT}-Ric_K(P)\right)x_t\\&+x_t'(A_K'P+PA_K)x_t.\end{aligned} \tag{9}$$

Proof: First, consider the following expansion of (7):

$$\mathscr{T}_KV(x)=x_t'\left(\int_0^T e^{A_K'\tau}Q_Ke^{A_K\tau}\,d\tau+e^{A_K'T}Pe^{A_KT}\right)x_t$$

Now, applying Lemma 1 to $e^{A_K'T}Pe^{A_KT}$ yields (8), the proof of the first part. Next, the differentiation of (7) leads to

$$\begin{aligned}\frac{d}{dt}\mathscr{T}_KV(x)&=x_{t+T}'Q_Kx_{t+T}-x_t'Q_Kx_t+x_{t+T}'(A_K'P+PA_K)x_{t+T}\\&=x_{t+T}'Ric_K(P)x_{t+T}-x_t'Ric_K(P)x_t+x_t'(A_K'P+PA_K)x_t.\end{aligned}$$

Therefore, the substitution of $x_{t+T}=e^{A_KT}x_t$ into the equation completes the proof of (9). ■

Using the operator $\mathscr{T}_K$, the exact value function $V_u(x)=x'P_Kx$ for a stabilizing $u=-Kx$ can be expressed as

$$\begin{aligned}V_u(x_t)&=\int_t^{t+T}x_\tau'Q_Kx_\tau\,d\tau+\int_{t+T}^\infty x_\tau'Q_Kx_\tau\,d\tau\\&=\mathscr{T}_KV_u(x).\end{aligned} \tag{10}$$

Similar expression is also possible for the equation of the Bellman's optimality principle [16]:

$$V_{u^*}(x_t)=\min_K\mathscr{T}_KV_{u^*}(x) \tag{11}$$

Algorithm 1: Generalized Policy Iteration

```
1: $i \leftarrow 0$
2: Initialize $P_0 \in \mathbb{M}^{n\times n}_{PS}$ and let $K_0 \leftarrow R^{-1}B'P_0$.
3: do {
4: Policy Evaluation:
       For a policy $K_i$, find $P_{i+1}$ which is an approximate of $P_{K_i}$
       satisfying $Ric_{K_i}(P_{K_i}) = 0$.
5: Policy Improvement:    $K_{i+1} \leftarrow R^{-1}B'P_{i+1}$
6: $i \leftarrow i+1$
7: Apply an exploration signal to excite the state $x$.
8: } until $\|P_i - P_{i-1}\| < \varepsilon$
```

where u^* and $V_{u^*}(x_t)$ are the optimal policy $u^* = -K^*x$ with the optimal gain $K^* \in \mathbb{M}^{m\times n}$ for LQR, and its corresponding optimal value function $V_{u^*}(x) = x'P_{K^*}x$ with the optimal index P_{K^*}, respectively. These two equations (10)–(11) are closely related to and actually the basis of the I-GPI.

From (11), one can obtain the optimal gain K^* as $K^* = R^{-1}B'P_{K^*}$ by the conventional optimal control arguments. Substituting this into the Lyapunov equation $Ric_{K^*}(P_{K^*}) = 0$ yields the algebraic Riccati equation (ARE) $Ric(P_{K^*}) = 0$, where the Riccati operator $Ric(P)$ is defined as $Ric(P) := A'P + PA - PBR^{-1}B'P + S$. Here, $Ric_K(P)$ and $Ric(P)$ satisfies

$$Ric(P) = Ric_K(P)\big|_{K=R^{-1}B'P}. \tag{12}$$

III. Generalized Policy Iteration

The usual GPI is shown in Algorithm 1 which consists of two successive steps—policy evaluation and policy improvement. These two steps are highly related to the equations $Ric_K(P_K) = 0$ and (12), respectively. In policy evaluation step (line 4), it tries to minimize the norm $\|Ric_{K_i}(P_{i+1})\|$ and as a result, gives an approximate solution P_{i+1} of P_{K_i} satisfying $Ric_{K_i}(P_{K_i}) = 0$. In policy improvement step (line 5), it updates K_{i+1} based on P_{i+1} to improve the policy K_{i+1} over K_i, that is, to achieve, for example, $\|Ric_{K_{i+1}}(P_{i+1})\| < \|Ric_{K_i}(P_i)\|$. This achievement indeed implies the improvement of the policy since the pair (P_i, K_i) always satisfies $Ric(P_i) = Ric_{K_i}(P_i)$ by (12) and $Ric(P_i) = 0$ holds whenever $P_i = P_{K^*}$. In line 7, some exploration signal is injected to the system (2) through u to hold the excitation condition which is necessary for the computation of P_i [5], [6], [9]–[13].

A. Integral GPI with DP Operator

The I-GPI is a class of the GPI methods, given in [11], to solve a given optimal control problem without knowing the system internal dynamics. This paper only focuses on the application to LQR and as a result, gives some mathematical properties including stability, monotonicity, and convergence. The basic operation of this algorithm is the one-step recursion at time $t \geq 0$, represented by

$$V_{i|j+1}(x_t) = \mathscr{T}_{K_i} V_{i|j}(x) \tag{13}$$

where $i \in \mathbb{N}$ is the iteration number, $j \in \mathbb{N}$ is the recursion index at i-th iteration, and $V_{i|j}(x)$ is a functional defined as $V_{i|j}(x) := x'P_{i|j}x$ for a matrix $P_{i|j} \in \mathbb{M}^{n\times n}$ indexed by (i,j). This one-step recursion (13) actually comes from the approximation of optimality principle (11), where $V_{u^*}(x)$ on the right hand side of (11) is replaced by $V_{i|j}$, assumed to be the most accurate approximate of V_{u^*} at (i,j)-th iteration. Now, the policy evaluation and improvement step of I-GPI applied to LQR can be derived from (13) as follows:

—- Algorithm 2: Integral GPI ———————————

Policy Evaluation: $V_{i+1}(x_t) = (\mathscr{T}_{K_i})^k V_i(x)$ (14)

Policy Improvement: $K_{i+1} = R^{-1}B'P_{i+1}$ (15)

where $V_i(x)$ is defined as $V_i(x) := x'P_ix$ for a indexed matrix $P_i \in \mathbb{M}^{n\times n}$ at i-th iteration. Here, the iteration horizon k represents the number of recursions (13) executed at each policy evaluation step. In [11], the authors mentioned that the policy evaluation (14) is a fixed point iteration, and proved the convergence to the exact value function V_{u^*} as $k \rightarrow \infty$, provided that the policy K_i is admissible. If $k = 1$, this I-GPI is actually the same as I-VI method [9], and as $k \rightarrow \infty$, I-GPI algorithm becomes the well-known I-PI [6], whenever (14) converges to a fixed point. This I-PI guarantees global stability and convergence [6] and is shown below:

—- Algorithm 3: Integral PI ——————————

Policy Evaluation: $V_{i+1}(x_t) = \mathscr{T}_{K_i} V_{i+1}(x)$ (16)

Policy Improvement: $K_{i+1} = R^{-1}B'P_{i+1}$ (17)

In this I-PI, policy evaluation step (16) exactly evaluates the value function $V_{i+1}(x_t)$ for the current policy K_i, which is same to the exact formula (10). In case of I-GPI with finite k, $V_{i+1}(x_t)$ can be an approximate of $V_{K_i}(x_t)$. Note that the error $|V_{i+1}(x_t) - V_{K_i}(x_t)|$ can be made arbitrarily small by adjusting k if $V_i(x_t)$ converges to $V_{K_i}(x_t)$ as $k \rightarrow \infty$. However, the large k introduces heavy computational burdens and hence, make the algorithm hard to implement in practice.

B. Policy Evaluation Step of I-GPI

We now mathematically explore policy evaluation step of I-GPI, and as a result, provide useful equivalent formula and convergence property, with the connection to the update horizon $\gamma > 0$ defined as a product of the iteration horizon k and time horizon T, that is, $\gamma := kT$. By using (13), the policy evaluation (14) of I-GPI can be represented as

Policy Evaluation: $V_{i|k}(x_t) = (\mathscr{T}_{K_i})^k V_{i|0}(x)$ (18)

where $V_{i|k}(x_t) := V_{i+1}(x_t)$ and $V_{i|0}(x_t) := V_i(x_t)$. For notational convenience, we define A_i as the matrix of i-th closed-loop system $A_i := A_{K_i}$ and $M_{i|j}$ as

$$M_{i|j} := \int_0^T e^{A_i'\tau} Ric_{K_i}(P_{i|j}) e^{A_i\tau}\, d\tau.$$

Consider the general k-th order recursive mapping

$$V_{i|j+k}(x_t) = (\mathscr{T}_{K_i})^k V_{i|j}(x). \tag{19}$$

If some properties regarding (19) are proven, then, they also holds for $V_{i|k}(x_t)$ and $V_{i|0}(x)$ satisfying (18) as a special case.

Here is the main theorem concerning k-th order recursive mapping (19) and its convergence:

Theorem 1: Consider the mapping (19) with the system (2). Then, for any $k \in \mathbb{N}$, $j \in \mathbb{Z}_+$, and $T > 0$, the mapping $V_{i|j+k}(x) = (\mathscr{T}_K)^k V_{i|j}(x)$ is equivalent to the followings:

$$1)\ Ric_{K_i}(P_{i|j+k}) = e^{A_i'(kT)} Ric_{K_i}(P_{i|j}) e^{A_i(kT)} \tag{20}$$

$$2)\ P_{i|j+k} = P_{i|j} + \int_0^{kT} e^{A_i'\tau} Ric_{K_i}(P_{i|j}) e^{A_i\tau}\, d\tau \tag{21}$$

$$\begin{aligned} 3)\ P_{i|j+k} = P_{i|j} - \left(Ric'_{K_i,P_{i|j}}\right)^{-1} \times \\ \left[Ric_{K_i}(P_{i|j}) - e^{A_i'(kT)} Ric_{K_i}(P_{i|j}) e^{A_i(kT)}\right] \end{aligned} \tag{22}$$

where $Ric'_{K_i,P_{i|j}}$ denotes the Frechet derivative of $Ric_{K_i}(P_{i|j})$ taken with respect to $P_{i|j}$. Moreover, if A_i is Hurwitz, then, $V_{i|j+k}(x_t)$ converges to the exact value function $V_{u_i}(x_t)$ as $\gamma\ (= kT) \to \infty$.

Proof: First, consider the one-step mapping $V_{i|j+1}(x) = \mathscr{T}_K V_{i|j}(x)$. Then, by (8) in Lemma 2, we have

$$P_{i|j+1} = P_{i|j} + \int_0^{T} e^{A_i'\tau} Ric_{K_i}(P_{i|j}) e^{A_i\tau}\, d\tau.$$

That is, $P_{i|j+1} = P_{i|j} + M_{i|j}$ in short. Now, by using this matrix equation, $Ric_{K_i}(P_{i|j+1})$ can be expressed in terms of $Ric_{K_i}(P_{i|j})$ as follows:

$$\begin{aligned} Ric_{K_i}(P_{i|j+1}) &= A_i'P_{i|j+1} + P_{i|j+1}A_i + K_i'RK_i + Q \\ &= Ric_{K_i}(P_{i|j}) + A_i'M_{i|j} + M_{i|j}A_i. \end{aligned} \tag{23}$$

where $A_i'M_{i|j} + M_{i|j}A_i$ can be rewritten as $A_i'M_{i|j} + M_{i|j}A_i = e^{A_i'T} Ric_{K_i}(P_{i|j}) e^{A_iT} - Ric_{K_i}(P_{i|j})$. Here, we used $A_i e^{A_i\tau} = e^{A_i\tau}A_i$ and Lemma 1. Substituting this into (23), we have

$$Ric_{K_i}(P_{i|j+1}) = e^{A_i'T} Ric_{K_i}(P_{i|j}) e^{A_iT}, \tag{24}$$

which is equivalent to the one-step mapping $V_{i|j+1}(x) = \mathscr{T}_{K_i} V_{i|j}(x)$. Therefore, (20) can be easily derived by recursively applying this relation as

$$\begin{aligned} Ric_{K_i}(P_{i|j+k}) &= e^{A_i'T} Ric_{K_i}(P_{i|j+k-1}) e^{A_iT} \\ &= \cdots = (e^{A_i'T})^k Ric_{K_i}(P_{i|j}) (e^{A_iT})^k. \end{aligned}$$

Next, we prove the equivalence between (21) and the mapping $V_{i|j+k}(x) = (\mathscr{T}_K)^k V_{i|j}(x)$. By employing $P_{i|j+1} = P_{i|j} + M_{i|j}$ to $P_{i|j+k}$ repetitively, one has

$$\begin{aligned} P_{i|j+k} &= P_{i|j+k-1} + M_{i|j+k-1} = P_{i|j+k-2} + M_{i|j+k-2} + M_{i|j+k-1} \\ &= \cdots = P_{i|j} + \sum_{l=0}^{k-1} M_{i|j+l}. \end{aligned}$$

Here, by (24), $M_{i|j+l}$ is

$$\begin{aligned} M_{i|j+l} &= \int_0^T e^{A_i'\tau} \cdot e^{A_i'T} Ric_{K_i}(P_{i|j+l-1}) e^{A_iT} \cdot e^{A_i\tau}\, d\tau \\ &= \int_T^{2T} e^{A_i'\tau} Ric_{K_i}(P_{i|j+l-1}) e^{A_i\tau}\, d\tau \\ &= \cdots = \int_{lT}^{(l+1)T} e^{A_i'\tau} Ric_{K_i}(P_{i|j}) e^{A_i\tau}\, d\tau. \end{aligned}$$

Therefore, one has

$$\sum_{l=0}^{k-1} M_{i|j+l} = \int_0^{kT} e^{A_i'\tau} Ric_{K_i}(P_{i|j}) e^{A_i\tau}\, d\tau,$$

which implies the equivalence between (21) and $V_{i|j+k}(x) = (\mathscr{T}_K)^k V_{i|j}(x)$. For the proof of (22), take the time derivative of the one-step mapping $V_{i|j+1}(x) = \mathscr{T}_{K_i} V_{i|j}(x)$ and employ (9) in Lemma 2. Then, one obtains

$$\begin{aligned} A_i'P_{i|j+1} + P_{i|j+1}A_i = e^{A_i'T} Ric_{K_i}(P_{i|j}) e^{A_iT} - Ric_{K_i}(P_{i|j}) \\ + A_i'P_{i|j} + P_{i|j}A_i \end{aligned} \tag{25}$$

which holds for all $j \in \mathbb{N} \cup \{0\}$. By iteratively applying (24)–(25) to $A_i'P_{i|j+k} + P_{i|j+k}A_i$, one obtains

$$\begin{aligned} & A_i'P_{i|j+k} + P_{i|j+k}A_i \\ &\quad = e^{A_i'T} Ric_{K_i}(P_{i|j+k-1}) e^{A_iT} - Ric_{K_i}(P_{i|j+k-1}) \\ &\qquad + A_i'P_{i|j+k-1} + P_{i|j+k-1}A_i \\ &\quad = (e^{A_i'T})^2 Ric_{K_i}(P_{i|j+k-2}) (e^{A_iT})^2 - Ric_{K_i}(P_{i|j+k-2}) \\ &\qquad + A_i'P_{i|j+k-2} + P_{i|j+k-2}A_i \\ &\quad\ \vdots \\ &\quad = (e^{A_i'T})^k Ric_{K_i}(P_{i|j}) (e^{A_iT})^k - Ric_{K_i}(P_{i|j}) + A_i'P_{i|j} + P_{i|j}A_i \end{aligned} \tag{26}$$

which is exactly same to the k-order recursive mapping (19) since we only employed (24)–(25) equivalent to the one step mapping $V_{i|j+1}(x) = \mathscr{T}_{K_i} V_{i|j}(x)$. Finally, rewriting (26) as

$$\begin{aligned} A_i'(P_{i|j+k} - P_{i|j}) + (P_{i|j+k} - P_{i|j})A_i \\ = e^{A_i'kT} Ric_{K_i}(P_{i|j}) e^{A_ikT} - Ric_{K_i}(P_{i|j}) \end{aligned} \tag{27}$$

yields (22) which is just another expression of (27) [9].

Now, consider the convergence of the mapping (19). Note that if the term $e^{A_i'kT} Ric_{K_i}(P_{i|j}) e^{A_ikT}$ converges to zero, then, (25) goes to the iteration

$$P_{i|j+k} = P_{i|j} - \left(Ric'_{K_i,P_{i|j}}\right)^{-1} Ric_{K_i}(P_{i|j}), \tag{28}$$

which is exactly the Kleinman's Newton method [10], [14]. This $e^{A_i'kT} Ric_{K_i}(P_{i|j}) e^{A_ikT} \to 0$ happens exactly when kT $(= \gamma)$ goes to infinity and A_i is Hurwitz. Therefore, whenever $\gamma \to \infty$, (25) becomes (28) which is equivalent to the PI [10]. Since policy evaluation of PI exactly calculates the value function $V_{u_i}(x) = x'P_{K_i}x$, the function $V_{i|j+k}(x) = x'P_{i|j+k}x$ goes to $V_{u_i}(x)$ as $\gamma \to \infty$, which completes the proof. ■

Note that the update term in (21) is the same as the term in (6) except the update horizon $\gamma\ (= kT)$ is finite and P_{K_i} is replaced with P_i. Obviously, this update term turns out to be zero when P_i equals to the exact value function P_{K_i}. By increasing γ, one can increase the update horizon of the integral. Furthermore, the increase of the update horizon γ can actually reduce the error $\|P_{i|j+k} - P_{K_i}\|$ when A_i is Hurwitz. Note that by the convergence argument, with Hurwiz matrix A_i, $\forall \varepsilon > 0$, $\exists \gamma^* \in \mathbb{N}$ such that $\forall \gamma \geq \gamma^*$, $\|P_{i|j+k} - P_{K_i}\| < \varepsilon$ holds. This can be also seen from (20) where sufficiently large γ makes the term

$e^{A_i kT}$ and thus $Ric_{K_i}(P_{i|j+k})$ arbitrarily small when A_i is Hurwitz. The Hurwitzness of A_i is certainly required for both the existence of P_{K_i} and the convergence of $P_{i|j+k}$ to P_{K_i}.

Remark 1: According to (20)–(22), all of the mappings $V_{i|j+k}(x) = (\mathscr{T}_{K_i})^k V_{i|j}(x)$ with the same update horizon $\gamma \in \mathbb{R}$ are actually all equivalent and have the same convergence speed, implying that the computational complexity due to the large iteration horizon k can be lessened by increasing the time horizon T for the same convergence speed.

Corollary 1: Consider the policy evaluation step (18)—$V_{i|k}(x_t) = (\mathscr{T}_{K_i})^k V_{i|0}(x)$. Then, it satisfies (20)–(22) with $j = 0$. Moreover, if A_i is Hurwitz, then, $V_{i|k}(x_t)$ converges to the exact value function $V_{u_i}(x_t)$ as $\gamma \to \infty$.

Remark 2: In [11], the convergence of $V_{i|k}(x_t)$ to $V_{u_i}(x_t)$ was proven with respect to $k \in \mathbb{N}$. On the other hand, Corollary 1 shows that the convergence result can be extended with respect to the update horizon $\gamma \in \mathbb{R}$.

C. Stability & Convergence of I-GPI Algorithm

Based on the results from Section III-B, we derive the local stability and convergence of the I-GPI algorithm (14)–(15). First, for notational convenience, define $\Phi_{(i,k)}$ and $M_{(i,k)}$ as

$$\Phi_{(i,k)} := \int_0^{kT} \|e^{A_i \tau}\|^2 \, d\tau, \tag{29}$$

$$M_{(i,k)} := \int_0^{kT} e^{A_i' \tau} Ric(P_i) e^{A_i \tau} \, d\tau. \tag{30}$$

In the policy improvement step (15), K_i is determined by $K_i = R^{-1}B'P_i$. Therefore, by incorporating this into the results in Section III-B, one obtains the following equivalent formulas:

Proposition 1: Consider the I-GPI algorithm (14)–(15). Then, it is equivalent to the following iterative forms:

1) $Ric(P_{i+1}) = e^{A_i'(kT)} Ric(P_i) e^{A_i(kT)} - M_{(i,k)} BR^{-1}B' M_{(i,k)}$ (31)

2) $P_{i+1} = P_i + \int_0^{kT} e^{A_i'\tau} Ric(P_i) e^{A_i \tau} \, d\tau$ (32)

3) $P_{i+1} = P_i - \left(Ric_{P_i}\right)^{-1} \left[Ric(P_i) - e^{A_i'(kT)} Ric(P_i) e^{A_i(kT)} \right].$ (33)

Proof: The equivalence to (32) and (33) can be easily derived by substituting $K_i = R^{-1}B'P_i$ into (21) and (22), respectively. For the proof of (31), consider $Ric(P_{i+1})$ with its expansion:

$$\begin{aligned} Ric(P_{i+1}) =& Ric(P_i) + A_i' M_{(i,k)} + M_{(i,k)} A_i \\ & - M_{(i,k)} BR^{-1}B' M_{(i,k)} \end{aligned} \tag{34}$$

where $A_i' M_{(i,k)} + M_{(i,k)} A_i$ can be represented as $A_i' M_{(i,k)} + M_{(i,k)} A_i = e^{A_i'(kT)} Ric(P_i) e^{A_i(kT)} - Ric(P_i)$ by Lemma 1 and $A_i e^{A_i T} = e^{A_i T} A_i$. By substituting this into (34), we have (31), which completes the proof. ■

In comparison to (20), the term $-M_i BR^{-1}B' M_i$ appears in the equation (31) which is caused by the policy improvement (15). The equation (31) plays a central role in the proof of convergence, and so do the two lemmas presented below:

Lemma 3: For any $(i,k) \in \mathbb{Z}_+^2$, $\Phi_{(i,k)}$ and $M_{(i,k)}$, defined by (29) and (30) respectively, satisfy the following inequality:

$$\|M_{(i,k)}\| \leq \Phi_{(i,k)} \|Ric(P_i)\|. \tag{35}$$

Proof: This can be directly proven by applying the property of matrix norm to $\|M_{(i,k)}\|$. ■

Lemma 4: Let P_{i+1} obtained by I-GPI (14)–(15) algorithm converges to P^*. Then, $P^* = P_{K^*}$ holds. That is, P^* is the LQ optimal solution satisfying $Ric(P^*) = 0$.

Proof: Assume $P_i \to P^*$. Then, taking the limit of (32) yields $0 = \lim_{P_i \to P^*} (P_{i+1} - P_i) = \lim_{P_i \to P^*} \int_0^{kT} e^{A_i' \tau} Ric(P_i) e^{A_i \tau} d\tau$, which implies $Ric(P^*) = 0$. Since $(A, B, Q^{1/2})$ is stabilizable and detectable, $Ric(P^*) = 0$ has a unique solution. Therefore, $P^* = P_{K^*}$ holds. ■

Now, we state the local stability and convergence of I-GPI. For a precise statement, we define linear and quadratic convergence as follows:

Definition 1: A sequence of matrix $\{P_i\}$ is said to converge to the solution P_{K^*}, linearly (resp. quadratically) in a set $\Omega \subset \overline{\Omega}$ if it locally converges to P_{K^*} in $\overline{\Omega}$, and have the property $\|Ric(P_{i+1})\| < \|Ric(P_i)\|$ (resp. $\|Ric(P_{i+1})\| < \|Ric(P_i)\|^2$) whenever $P_i \in \Omega$. We also say that $\{P_i\}$ linearly (resp. quadratically) converges to P_{K^*} if $\Omega = \overline{\Omega}$.

Theorem 2: Consider the I-GPI algorithm (14)–(15) with the system (2) and define the bounds C_i, D_i, and E_i as $C_i := (2\left\|BR^{-1}B'\right\| \|Y_i\| \Phi_{(i,k)})^{-1}$,

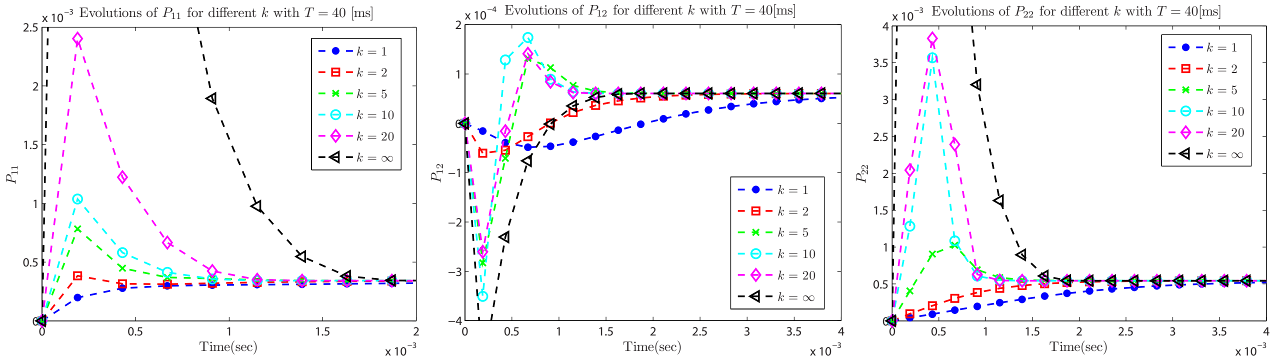
$$D_i := \frac{1 - \|e^{A_i(kT)}\|^2}{\|BR^{-1}B'\| \Phi_{(i,k)}^2}, \quad E_i := \frac{\|e^{A_i(kT)}\|^2}{1 - \|BR^{-1}B'\| \Phi_{(i,k)}^2},$$

respectively. Suppose $\overline{\Omega}_{D_i}$ and Ω_{E_i} be the sets defined as

$$\begin{aligned} \overline{\Omega}_{D_i} &:= \{P \in \mathbb{M}^{n \times n} : \|Ric(P)\| < D_i\}, \\ \Omega_{E_i} &:= \{P \in \mathbb{M}^{n \times n} : E_i < \|Ric(P)\| < 1\}, \end{aligned}$$

respectively. Then, for all $i \in \mathbb{N} \cup \{0\}$,

1. (**stability**) if A_i Hurwitz, so is A_{i+1} when $\|Ric(P_i)\| \leq C_i$;
2. (1**st-order monotonicity**) if $P_i \in \overline{\Omega}_{D_i}$ at i-th iteration, then, P_{i+1} satisfies $\|Ric(P_{i+1})\| < \|Ric(P_i)\|$;
3. (2**nd-order monotonicity**) if $\left\|BR^{-1}B'\right\| \Phi_{(i,k)}^2 \neq 1$ and $E_i < \|Ric(P_i)\|$ is satisfied at i-th iteration, then, P_{i+1} satisfies $\|Ric(P_{i+1})\| < \|Ric(P_i)\|^2$;



(a)

(b)

(c)

Fig. 1. The evolutions of critic parameters P_i with $T = 40$[ms] for various k_i's–(a) P_{11}, (b) P_{12}, and (c) P_{22}.

4. (**linear convergence & quadratic decreasing**) if $P_i \in \overline{\Omega}_{D_i}$ for all $i \in \mathbb{N} \cup \{0\}$, then, P_i linearly converges to P_{K^*}; moreover, if D_i is larger than 1 ($1 < D_i$), then, $\Omega_{E_i} \neq \varnothing$ and $\Omega_{E_i} \subset \overline{\Omega}_{D_i}$ holds; in this case, P_i converges to P_{K^*}, quadratically in Ω_{E_i}.
5. (**quadratic convergence of policy iteration**) if $\{P_i\}$ is generated by policy iteration (16)–(17), then it quadratically converges to P_{K^*} whenever $P_0 \in \Omega_0$ ($\Omega_0 = \Omega_{E_i}|_{E_i=0}$) and A_0 is Hurwitz.

Proof: For the proof of the stability part, follow the same procedure given in [13], with $M_{(i,k)}$ defined in (30) (see the proof of Theorem 1 [13]). For the proof of the monotonicity, take the matrix norm $\|\cdot\|$ of (31) and employ Lemma 3 and the properties of the norm as follows:

$$\begin{aligned} \|Ric(P_{i+1})\| &\le \|e^{A_i(kT)}\|^2 \cdot \|Ric(P_i)\| + \|M_i\|^2 \cdot \|BR^{-1}B'\| \\ &\le \|e^{A_i(kT)}\|^2 \|Ric(P_i)\| + \Phi_k^2 \|BR^{-1}B'\| \|Ric(P_i)\|^2. \end{aligned} \tag{36}$$

By applying $\|Ric(P_i)\| < D_i$ to (36), one proves the 1st-order monotonicity $\|Ric(P_{i+1})\| < \|Ric(P_i)\|$. Next, assume $\|Ric(P_i)\| < E_i$ and rearrange (36) as

$$\|Ric(P_{i+1})\| \le \left(\frac{\|e^{A_i(kT)}\|^2}{\|Ric(P_i)\|} + \Phi_k^2 \|BR^{-1}B'\| \right) \|Ric(P_i)\|^2.$$

Then, by inverting $\|Ric(P_i)\| < E_i$ and applying it to the above inequality $\|Ric(P_{i+1})\| < \|Ric(P_i)\|^2$ can be proven, the 2nd-order monotonicity.

In the sequel, we will focus on the convergence of P_i. if $\|Ric(P_i)\| < D_i$ holds for all $i \in \mathbb{N} \cup \{0\}$, then by 1st-order monotonicity $\|Ric(P_{i+1})\| < \|Ric(P_i)\|$ and lower boundedness of $\|Ric(P_i)\|$ by zero, $\{\|Ric(P_i)\|\}$ converges, and so does $\{Ric(P_i)\}$ with this topology. Therefore, by Lemma 4, we conclude that P_i linearly converges to P_{K^*} whenever $P_i \in \overline{\Omega}_{D_i}$ for all $i \in \mathbb{N} \cup \{0\}$.

For the proof of quadratic decreasing, suppose $D_i > 1$ and rearrange the inequality. Then, one obtains $1 - \|BR^{-1}B'\| \cdot \Phi_{(i,k)}^2 > 0$ and $E_i < 1$. Therefore, $\Omega_{E_i} \neq \varnothing$ is valid, and thus, by 2nd-order monotonicity, $\|Ric(P_{i+1})\| \le \|Ric(P_i)\|^2$ holds whenever $P_i \in \Omega_{E_i}$. In this case, $\Omega_{E_i} \subset \overline{\Omega}_{D_i}$ is obvious by $E_i < 1 < D_i$. Note that $P_i \to P_{K^*}$ if $P_i \in \overline{\Omega}_{D_i}$ $\forall i \in \mathbb{N} \cup \{0\}$ by linear convergence argument. Then, it is obvious that P_i converges to P_{K^*}, quadratically in Ω_{E_i} ($\subset \overline{\Omega}_{D_i}$).

Now, note that if A_i is Hurwitz and $\gamma \to \infty$, E_i goes to zero and I-GPI (14)–(15) becomes the I-PI (16)–(17). Since I-PI yields Hurwitz A_i when A_0 is Hurwitz, one can assume A_i is Hurwitz without loss of generality. Now, considering the metric the metric $d(\Omega_{E_i}, \Omega_{E_j}) = |E_i - E_j|$ on $\{\Omega_E : 0 \le E < 1\}$, one can see $\lim_{\gamma \to \infty} \Omega_{E_i} = \Omega_0$. Therefore, by quadratic decreasing, $\{P_i\}$ generated by (16)–(17) quadratically converges to P_{K^*} whenever $P_0 \in \Omega_0$, which completes the proof. ■

Remark 3: Although the monotonicity and convergence were proven independently of the stability of A_i, it is actually related to the existence of $P \in \mathbb{M}^{n \times n}$ such that $0 < \|Ric(P)\| < D_i$ ($P \in \overline{\Omega}_{D_i}$) holds. Note that to satisfy $0 < D_i$ for any $k \in \mathbb{N}$ and any $T > 0$, A_i should be at least Hurwitz, which is the connection between C_i and D_i.

Remark 4: As is well-known in the literature [14], I-PI (16)–(17) guarantees quadratic convergence in the vicinity of $Ric(P_{K^*}) = 0$. In this article, a concrete set Ω_0 around P_{K^*} is provided in which the I-PI (16)–(17) achieves quadratic convergence.

IV. Numerical Results

To verify the claims raised in Section III, the I-GPI (14)–(15) is simulated with the following step-down converter model for various $k \in \mathbb{N}$:

$$A = \begin{bmatrix} 0 & -1/L \\ 1/C & -1/R_oC \end{bmatrix}, \; B = \begin{bmatrix} V_g/L \\ 0. \end{bmatrix}$$

where the parameters were set to $L = 200$ [μH], $C = 200$ [μF], $R_o = 25$ [Ω], and $V_g = 24$ [V], and the performance index (3) with $S = \text{diag}\{5, 1\}$ and $R = 300$ is considered. With these settings, P_{K^*} associated with the optimal value function $V_{u^*}(x)$ can be evaluated as

$$P_{K^*} = \begin{bmatrix} 0.3418 & 0.0606 \\ 0.0606 & 0.5431 \end{bmatrix} \times 10^{-3}. \tag{37}$$

Furthermore, for the implementation of I-GPI (14)–(15), the batch least squares, already used in [6], [9]–[13], is adopted for the calculation of P_{i+1} in policy evaluation where

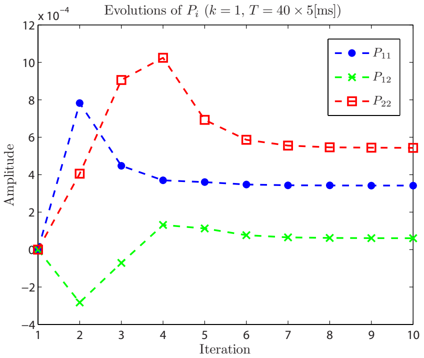


Fig. 2. The evolutions of P_i when $k = 1$ and $T = 40 \times 5$ [ms].

TABLE I
THE EVALUATIONS OF C_i, D_i, AND E_i THROUGH SIMULATIONS

γ	C_i		D_i		E_i	
	$\min C_i$	C^*	$\min D_i$	D^*	$\min E_i$	E^*
1	0.94	0.94	0.06	0.16	1.07	1.07
2	0.51	0.23	0.20	0.51	1.27	1.27
5	0.67	0.67	2.54	2.54	0.69	0.73
10	1.34	1.34	21.27	21.27	0.19	0.20
20	2.68	2.68	104.39	104.4	0.01	0.02
∞	13.40	13.40	2651.1	2651.1	0.00	0.00

5 number of data samples (x_t, u_t) are collected a iteration. After each policy improvement, an exploration signal $w(t) = 10^{-2} \sin 2\pi f$ with $f = 50$ [kHz] is applied during one period T through u to prevent x_t from being stationary.

Fig. 1 illustrates the trajectories of the convergent parameters P_{11}, P_{12}, and P_{22} for various $k \in \mathbb{N}$ with the time horizon $T = 40$ [ms] and initial policy $u_0 \equiv 0$. As can be seen from the figures, the convergence speeds of P_i are lowest when $k = 1$ (I-VI case). As $k \in \mathbb{N}$ is increased, the convergence to P_{K^*} tends to be achieved more rapidly than the case $k = 1$, but introduces much higher overshoots at $i = 1$ especially when $k \geq 10$. Note that the largest overshoots appear when $k = \infty$ (I-PI case), and in this case, the convergence speeds do not seem to be significantly improved in comparison to the other case $(k < \infty)$. Therefore, the choice of suitable k can be a main issue which achieves an appropriate trade-off between the convergence speeds and the degree of the overshoots.

Fig. 2 describes the evolutions of P_i when $u_0 \equiv 0$, $k = 1$, and $T = 40 \times 5$ [ms]. Comparing Fig. 2 with the case $k = 5$ and $T = 40$ [ms] in Fig. 1 in iteration domain, one can see that both P_i's evaluated by (14) with the same update horizon $\gamma = 1 \times 5 \times 40 = 200$ [ms] are exactly same to each other. This verifies the claims from Theorem 1 and Proposition 1—I-GPI algorithms (14)–(15) with the same update horizon γ are all equivalent.

To investigate how much the bounds C_i, D_i, and E_i are conservative, additional simulations are carried out for the different γ. Since Theorem 2 provides the *local* stability and convergence in the vicinity of $Ric(P_{K^*}) = 0$, we employ $K_0 = R^{-1}B'P_0$ as the initial policy u_0, where P_0 is the solution of ARE with A replaced by the nominal matrix A_{nom}. Here, A_{nom} is composed of $L_{nom} = 250$ [uH], $C_{nom} = 285.72$ [uF], and $R_{o,\,nom} = 13.46$ [Ω]. Table I shows some bounds for different γ, where minimum is taken over the whole iterations, and C^*, D^*, and E^* denote the values of C_i, D_i, and E_i after P_i converges. Through the result, one can see that E_i converges to 0 as $\gamma \to \infty$, enlarging the region of quadratic convergence. Also note that the larger γ, the larger areas of stability and convergence bounds $\bar{\Omega}_{D_i}$ can be achieved.

V. CONCLUDING REMARKS

In summary, we have provided the various equivalent formula of GPI with respect to the update horizon $\gamma\ (= kT)$, which revealed the relationships between the time horizon T and the computational complexity due to k. The criteria regarding local stability and convergence were also provided for I-GPI. However, I-PI, the special case of I-GPI, actually guarantees the *global* stability and convergence [10], which implies there would be less conservative bounds than those in Theorem 2. Therefore, the future works would be to find such bounds or global stability conditions of I-GPI, make research on the relations of discount factor and learning rate [13] to I-GPI, and extend the results to the general nonlinear systems.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning–An Introduction*, MIT Press, Cambridge, Massachussetts, 1998.

[2] W. B. Powell *Approximate Dynamic Programming–Solving the Curse of Dimensionality*, Wiley-Interscience, 2007.

[3] J. Si, A. G. Barto, W. B. Powell, and D. Wunsch, *Handbook of Learning and Approximate Dynamic Programming*, Wiley-IEEE Press, 2004.

[4] F. Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: an introduction," *IEEE Computational Magazine*, vol. 4, no. 2, pp. 39–47.

[5] S. J. Bradtke and B. E. Ydstie, "Adaptive linear quadratic control using policy iteration," *Proc. American Control Conference*, Baltimore, Maryland, pp. 3475–3479, 1994.

[6] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[7] K. Doya, "Reinforcement learning in continuous-time and space," *Neural Computation 12*, pp. 219–245, 2000.

[8] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. saeks, "Adaptive Dynamic Programming," *IEEE Trans. Systems, Mans and Cybernetics*, vol. 32, no. 2, pp. 140–153, 2002.

[9] D. Vrabie, M. Abu-Khalaf, F. L. Lewis, and Y. Wang, "Continuous-time ADP for linear systems with partially unknown dynamics," *Proc. the IEEE Int. Symp. Approximate Dynamic Programming and Reinforcement Learning*, pp. 247–253, 2007.

[10] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[11] D. Vrabie and F. L. Lewis, "Generalized policy iteration for continuous-time systems," *Proc. Int. Joint Conf. Neural Networks*, Alnata, GA, USA, pp. 3224–3231, 2010.

[12] J. Y. Lee , J. B. Park, and Y. H. Choi, "Model-free approximate dynamic programming for continuous-time linear systems," *Proc. Decision and Control, held joinly with Chinese Control Conf.*, Shanghai, China, pp. 5009–5014, 2009.

[13] J. Y. Lee, J. B. Park, and Y. H. Choi, "A generalized value iteration scheme for continuous-time linear systems," *Proc. Decision and Control*, Atlanta, GA, USA, pp. 4637–4642, 2010.

[14] D. Kleinman, "On the iterative technique for Riccati equation computations," *IEEE Trans. Automatic Control*, vol. 13, no. 1, pp. 114–115, 1968.

[15] H. K. Khalil, *Nonlinear Systems*, Prentice Hall, 2002.

[16] D. E. Kirk, *Linear Optimal Control: An Introduction*, Dover Publications New York, 2004.