On Integral Generalized Policy Iteration for Continuous-Time Linear Quadratic Regulations ⋆

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

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

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

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.

Key words: LQR, generalized policy iteration, reinforcement learning, adaptive control, optimization under uncertainties

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 van Nunen (1976) and Puterman & Shin (1978) 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,

⋆ 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. This paper was not presented at any IFAC.

* Corresponding author. Tel.: +82 2 2123 2773; fax: +82 2 362 4539.

Email addresses: jyounglee@yonsei.ac.kr (Jae Young Lee), jbpark@yonsei.ac.kr (Jin Bae Park), yhchoi@kyonggi.ac.kr (Yoon Ho Choi).

Liu, Wei, Zhao, & Jin, 2012; Webos, 1992; Zhang, Huang, & Lewis, 2009) and recently, continuous-time (CT) domain (Bhasin, Kamalapurkar, Johnson, Vamvoudakis, Lewis, & Dixon, 2013; Doya, 2000; Hanselmann, Noakes, & Zaknich, 2007; Lee, Park, & Choi, 2010, 2012; Vamvoudakis & Lewis, 2010; Vrabie, 2009; Vrabie & Lewis, 2009).Wang, Zhang, & Liu (2009) and Lewis & Vrabie (2009) performed recent 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 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 2^{nd}-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 & 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 & 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 2^{nd}-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 & Lewis (2009) proved that, under an admissible policy, the value function approx-

imated 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 & 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 & 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 to $\hbar$, where the infinitesimal version of I-GPI ($\hbar \to 0$) is at one extreme tip, and I-PI ($\hbar \to \infty$) belongs to the other extreme tip of the spectrum, as shown in Fig. 1.

All of these analytical results are derived based on the positive definiteness property of the evaluated value functions and the matrix iterative formulas equivalent to I-GPI, also provided in this paper. Data-driven least squares (LS) implementation methods of I-GPI are also proposed, including the adaptive methods for determining k at each step without violating the presented matrix inequalities. All of those implementation methods are based on a matrix recursion, where the multiplicative matrix terms are structured from the measured data that is sufficiently persistently exciting. Finally, several numerical simulations are carried out to verify and further investigate the individual mathematical properties and LS implementation methods of I-GPI.

This paper is organized as follows. In Section 2, the target

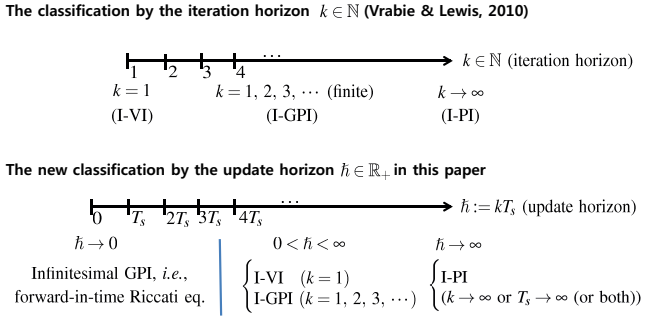


Fig. 1. The classifications of I-RL algorithms.

CT LQR problem is formulated with the associated matrix operators. In Section 3, we introduce the main I-GPI algorithm represented by the update horizon $\hbar$ in LQR frameworks and show the equivalences of I-GPIs that have the same $\hbar$. Section 4 is devoted to proving the properties of the policy evaluation recursion of each step of I-GPI such as positive definiteness and monotone convergence in the limit $\hbar \to \infty$. The stability and monotone convergence properties of I-GPI including those in PI- and VI-modes are shown in Section 5 with detailed discussions. The new classification of I-RL with respect to $\hbar$ is also established in this section. Section 6 illustrates the (adaptive) data-driven LS implementation methods in LQR frameworks. Finally, the numerical simulation results for load frequency control systems are illustrated in Section 7 and conclusions follow in Section 8.

Notations: The mathematical symbols used in this paper are summarized as follows. $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$ denotes the set of nonnegative real numbers; $\mathbb{Z}_+ := \mathbb{N} \cup \{0\}$ is the set of nonnegative integers; the set of all $m \times n$ constant real matrices is denoted by $\mathbb{M}^{m\times n}$. For a matrix $X \in \mathbb{M}^{n\times n}$, $\lambda_i(X)$ denotes the i-th eigenvalue of X with the decreasing order $\mathrm{Re}\lambda_n(X) \leq \mathrm{Re}\lambda_{n-1}(X) \leq \cdots \leq \mathrm{Re}\lambda_1(X)$. We also denote $\|Z\|$ for a matrix $Z \in \mathbb{M}^{n\times m}$ the spectral norm of Z, *i.e.*, $\|Z\| := \sqrt{\lambda_1(Z^TZ)}$.

2 LQR problems and Lyapunov/Riccati operators

In this section, we introduce and briefly discuss LQR problems and two related operators $\mathscr{L}(K,P)$ and $\mathscr{R}(P)$, named Lyapunov and Riccati operators, respectively, with special focus on the value function $V_u(x)$ for a stabilizing policy and the optimal solution $(u^*, V_{u^*}(x))$. Consider the CT linear system ($t \geq 0$):

$$\dot{x}_t = Ax_t + Bu_t, \tag{1}$$

for the state $x_t \in \mathbb{R}^n$, the 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 performance index

$$V_u(x_t,t) = \int_t^\infty x_\tau^T S x_\tau + u_\tau^T R u_\tau \, d\tau$$

where $S \in \mathbb{M}^{n\times n}$ and $R \in \mathbb{M}^{m\times m}$ are positive semi-definite and positive definite matrices, respectively. Throughout the paper, $u(t)$, u_t, and simply u will be used interchangeably for the input of the system (1).

Let $u = -Kx$, or simply K, be any linear policy for the system (1) and A_K be its corresponding closed loop matrix $A - BK$. Define Q_K for a policy K as $Q_K := S + K^T R K$ for simplicity. Then, $V_u(x_t, t)$ can be represented in terms of Q_K as

$$V_u(x_t,t) = x_t^T \left(\int_t^\infty e^{A_K^T(\tau - t)} Q_K e^{A_K(\tau - t)}\, d\tau \right) x_t = x_t^T P_K x_t,$$

where P_K is defined as

$$P_K := \int_0^\infty e^{A_K^T \tau} Q_K e^{A_K \tau}\, d\tau. \tag{2}$$

If $u = -Kx$ is stabilizing [1], then the value function $V_u(x) = x^T P_K x$ is always finite and P_K is the unique positive semi-definite solution of the Lyapunov equation $\mathscr{L}(K, P_K) = 0$, where the Lyapunov operator $\mathscr{L}(K,P)$ is defined as

$$\mathscr{L}(K,P) := A_K^T P + P A_K + Q_K. \tag{3}$$

Here, $\mathscr{L}(K, P_K) = 0$ can be easily verified by substituting (2) into (3) and using standard calculus. Note that the optimal policy $u^* = -K^* x$ and its corresponding optimal value function $V_{u*}(x) = x^T P_{K^*} x$ also satisfy the Lyapunov equation $\mathscr{L}(K^*, P_{K^*}) = 0$ and the optimal policy is given by $K^* = R^{-1} B^T P_{K^*}$ by the conventional optimal control theory. Substituting the optimal gain $K^* = R^{-1} B^T P_{K^*}$ into the Lyapunov equation and rearranging the equation, we obtain the standard ARE $\mathscr{R}(P_{K^*}) = 0$, where the Riccati operator $\mathscr{R}(P)$ is defined as

$$\mathscr{R}(P) := A^T P + PA - PBR^{-1}B^T P + S$$

and satisfies $\mathscr{R}(P) = \mathscr{L}(K,P)|_{K = R^{-1}B^T P}$. The objective of the GPI methods is to find the optimal solution (K^*, P_{K^*}) in online fashion by an iterative procedure of solving the Lyapunov equation $\mathscr{L}(K, P_K) = 0$ and updating the control policy K. For the existence of the unique stabilizing solution P_{K^*}, we assume that

Assumption 1 *The triple $(A, B, S^{1/2})$ is stabilizable and detectable.*

For a given matrix $P \in \mathbb{M}^{n\times n}$, we also define the P-dependent control gain matrix K_P as $K_P := R^{-1} B^T P$ for notational convenience. Using this notation, we have $\mathscr{R}(P) = \mathscr{L}(K_P, P)$ from $\mathscr{R}(P) = \mathscr{L}(K,P)|_{K = R^{-1}B^T P}$ and the following lemma can be obtained which will be extensively used in the analysis of the I-GPI algorithm.

[1] In this paper, a policy $u = -Kx$ is said to be a stabilizing policy (or simply, stabilizing) if A_K is Hurwitz.

Lemma 1 *For any $P, \Phi \in \mathbb{M}^{n\times n}$ and $K \in \mathbb{M}^{m\times n}$, the operators $\mathscr{L}(\cdot,\cdot)$ and $\mathscr{R}(\cdot)$ satisfy the followings:*

- $\mathscr{L}(K,P) - \mathscr{L}(K,\Phi) = A_K^T(P - \Phi) + (P - \Phi)A_K,$ (4)
- $\mathscr{L}(K,P) = \mathscr{R}(P) + (K - K_P)^T R (K - K_P).$ (5)

Proof. (4) can be easily verified by substituting (3). For the proof of (5), note that $\mathscr{R}(P)$ can be represented in terms of K_P as

$$\mathscr{R}(P) = A^T P + PA - K_P^T R K_P + S,$$

and that $(K - K_P)^T R (K - K_P) = K^T R K - K_P^T R K - K^T R K_P + K_P^T R K_P$. Then, the proof is completed by substituting these and (3) into (5) and rearranging the equation. □

3 Integral generalized policy iteration and its equivalence

In this section, we present and re-derive the I-GPI algorithms (Vrabie & Lewis, 2009) in LQR frameworks, with discussions of dynamic programming (DP) operators, optimality principles, and their generalizations. Then we show the equivalence of all the I-GPI algorithms that have the same update horizon $\hbar$.

Regarding the system dynamics (1), the DP operator $\mathscr{T}_K^{T_s} : X \to X$ is defined on the space X of continuous functionals $V(x) : \mathbb{R}^n \to \mathbb{R}$, at fixed time $t \geq 0$, as

$$\mathscr{T}_K^{T_s} V(x_t) := \int_t^{t+T_s} x_\tau^T Q_K x_\tau\, d\tau + V(x_{t+T_s}), \tag{6}$$

where the trajectories of x_t are generated by the system (1) with a given control $u = -Kx$. From this, the generalized DP operator $(\mathscr{T}_K^{T_s})^k$ can be recursively defined for $k \in \mathbb{Z}_+$ and $T_s \in \mathbb{R}_+$ as

$$\begin{cases} (\mathscr{T}_K^{T_s})^0 V(x_t) := V(x_t), \\ (\mathscr{T}_K^{T_s})^{k+1} V(x_t) := (\mathscr{T}_K^{T_s})^k [\mathscr{T}_K^{T_s} V(x_t)] \end{cases}$$

at fixed time $t \geq 0$. Here, k is referred to as the number of the DP operation $\mathscr{T}_K^{T_s}[\cdot]$. In this paper, we call $k \in \mathbb{Z}_+$ and $T_s \in \mathbb{R}_+$ 'iteration horizon' and 'time horizon', respectively. This DP operator simplifies the mathematical statements related to the integral temporal difference (I-TD), optimality principle, and the I-GPI algorithm. Indeed, the value function $V_u(x) = x^T P_K x$ for a stabilizing policy $u = -Kx$ can be expressed as the following I-TD form:

$$\begin{aligned} V_u(x_t) &= \int_t^{t+T_s} x_\tau^T Q_K x_\tau\, d\tau + \underbrace{\int_{t+T_s}^\infty x_\tau^T Q_K x_\tau\, d\tau}_{= V_u(x_{t+T_s})} \\ &= \mathscr{T}_K^{T_s} V_u(x_t). \end{aligned} \tag{7}$$

A similar expression is also possible for the equation of Bellman's optimality principle (Lewis & Vrabie, 2009):

$$V_{u^*}(x_t) = \min_K \mathscr{T}_K^{T_s} V_{u^*}(x_t). \tag{8}$$

By using the generalized DP operator, these simple expressions (7)–(8) can be easily extended to the general case with respect to the update horizon $\hbar \in \mathbb{R}_+$ defined as $\hbar := kT_s$ in this paper.

Theorem 1 *The DP operator $\mathscr{T}_K^{T_s}$ and its generalized operator $(\mathscr{T}_K^{T_s})^k$ for a continuous functional $V(x)$ satisfy*

$$(\mathscr{T}_K^{T_s})^k V(x_t) = \mathscr{T}_K^{\hbar} V(x_t). \tag{9}$$

Proof. Consider the sequence $\{W_i(x_t)\}_{i=0}^k$ of continuous functionals which is defined by $W_0(x_t) := V(x_t)$ and $W_i(x_t) := \mathscr{T}_K^{T_s} W_{i-1}(x_t)$ for $i = 1, 2, 3, \cdots, k$. Then, obviously, $(\mathscr{T}_K^{T_s})^k V(x_t) = W_k(x_t)$ holds and thereby, one has

$$\begin{aligned}
(\mathscr{T}_K^{T_s})^k V(x_t) &= \int_t^{t+T_s} x_\tau^T Q_K x_\tau \, d\tau + \mathscr{T}_K^{T_s}\big[\underbrace{(\mathscr{T}_K^{T_s})^{k-2} V(x_{t+T_s})}_{=W_{k-2}(x_{t+T_s})}\big] \\
&= \int_t^{t+2T_s} x_\tau^T Q_K x_\tau \, d\tau + \mathscr{T}_K^{T_s} W_{k-3}(x_{t+2T_s}) \\
&\vdots \\
&= \int_t^{t+(k-1)T_s} x_\tau^T Q_K x_\tau \, d\tau + \mathscr{T}_K^{T_s} W_0(x_{t+(k-1)T_s}) \\
&= \int_t^{t+kT_s} x_\tau^T Q_K x_\tau \, d\tau + W_0(x_{t+kT_s}) \\
&= \mathscr{T}_K^{kT_s} V(x_t),
\end{aligned}$$

which completes the proof. □

By Theorem 1, the I-TD formula (7) and the optimality equation (8) can be easily extended with the generalized DP operator $(\mathscr{T}_K^{T_s})^k$ as

$$V_u(x_t) = (\mathscr{T}_K^{T_s})^k V_u(x_t), \tag{10}$$

$$V_{u^*}(x_t) = \min_K \big[(\mathscr{T}_K^{T_s})^k V_{u^*}(x_t)\big], \tag{11}$$

both of which, together with (7) and (8), are closely related to the I-GPI. Assuming, at i-th iteration, V_i to be the most accurate approximation of V_{u^*} on the right hand side of (11), one obtains the I-GPI algorithm shown in Algorithm 1 consisting of two main successive steps—approximate policy evaluation (lines 6–10) and policy improvement (line 11). At each i-th approximate policy evaluation step, the next value function $V_{i+1}(x)$, defined as $V_{i+1}(x) := x^T P_{i+1} x$ for an

Algorithm 1: I-GPI for LQR

1: $i \leftarrow 0$
2: Initialize $P_0 \in \mathbb{M}^{n\times n}$.
3: Set an initial policy $u_0 = -K_0 x$ not necessarily stabilizing.
4: **do** {
5: Apply the current policy $u_i(t) = -K_i x(t)$ to the system (1).

Approximate policy evaluation: $V_{i+1}(x_t) = (\mathscr{T}_{K_i}^{T_s})^k V_i(x_t)$
6: $P_{i|0} \leftarrow P_i$
7: **for** $j = 0$ **to** $k-1$,
8: find $V_{i|j+1}(x)$ $\big(= x^T P_{i|j+1} x\big)$ by solving (12).
9: **end**
10: $P_{i+1} \leftarrow P_{i|k}$

Policy improvement: $K_{i+1} = R^{-1} B^T P_{i+1}$
11: $K_{i+1} \leftarrow R^{-1} B^T P_{i+1}$
12: $i \leftarrow i+1$
13: Apply an **exploration** signal to excite the state x.
14: } **until** $\|P_i - P_{i-1}\| < \varepsilon$.

indexed matrix $P_{i+1} \in \mathbb{M}^{n\times n}$, is obtained by performing the basic one-step DP recursion at time $t \geq 0$

$$V_{i|j+1}(x_t) = \mathscr{T}_{K_i}^{T_s} V_{i|j}(x_t) \tag{12}$$

k-times (lines 6–10) for the applied current policy $u_i = -K_i x$ (line 5), where k is the iteration horizon representing the number of recursions (12), $j \in \{0, 1, 2, \cdots, k-1\}$ is the recursion index at the i-th iteration, and $V_{i|j}(x)$ is a functional defined as $V_{i|j}(x) := x^T P_{i|j} x$ for a matrix $P_{i|j} \in \mathbb{M}^{n\times n}$ indexed by (i, j). Then, the next policy K_{i+1} is updated at each i-th policy improvement step (line 11) based on P_{i+1}. In line 13, some exploration signal is injected into the system (1) through u to hold the excitation condition, which is necessary for the computation of P_i (Lee *et al.*, 2011; Lewis & Vrabie, 2009) and will be discussed in Section 6. This whole procedure continues until the value function matrix P_i converges (line 14).

Notice K_i converges whenever P_i does. In addition, the next lemma states that the convergent point corresponds to the optimal solution (u^*, V_{u^*}). This lemma will be used to finalize the proofs of monotone convergence in Section 5.

Lemma 2 *Consider the sequences $\{P_i\}_{i=0}^\infty$ and $\{K_i\}_{i=0}^\infty$ generated by the I-GPI (Algorithm 1) and let $\{P_i\}$ be a convergent sequence. Then, under Assumption 1, P_i and K_i converge to the optimal solutions P_{K^*} and K^*, respectively.*

Proof. See Appendix. □

Unlike the I-GPI given by Vrabie & Lewis (2009), Algorithm 1 does not assume that the initial policy is stabilizing (line 3); all the other parts of both I-GPIs are the same in LQR frameworks. In case of a stabilizing policy K_i at i-th iteration,

Algorithm 2: I-PI for LQR

1: $i \leftarrow 0$
2: Initialize $P_0 \in \mathbb{M}^{n\times n}$ (to zero).
3: Set an initial stabilizing policy $u_0 = -K_0 x$.
4: **do** {
5: Apply the current policy $u_i(t) = -K_i x(t)$ to the system (1).

Policy evaluation: $V_{u_i}(x_t) = \mathscr{T}_{K_i}^{T_p} V_{u_i}(x_t)$

6: Find $V_{i+1}(x)$ $(= x^T P_{i+1} x)$, which is equal to the exact one $V_{u_i}(x)$ $(= x^T P_{K_i} x)$, by solving $V_{i+1}(x_t) = \mathscr{T}_{K_i}^{T_p} V_{i+1}(x_t)$.

Policy improvement: $K_{i+1} = R^{-1}B^T P_{K_i}$

7: $K_{i+1} \leftarrow R^{-1}B^T P_{i+1}$
8: $i \leftarrow i+1$
9: Apply an **exploration** signal to excite the state x.
10: } **until** $\|P_i - P_{i-1}\| < \varepsilon$.

- the policy evaluation step (line 5–9) attempts to approximate P_{K_i} satisfying $\mathscr{L}(K_i, P_{K_i}) = 0$, and as a result, gives an approximate solution P_{i+1} of P_{K_i};
- the policy improvement step (line 10) 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, $0 \le P_{K_{i+1}} \le P_{K_i}$.

Actually, Algorithm 1 with $k = \infty$ under an initial stabilizing policy is the same as I-PI (Algorithm 2); Algorithm 1 with $k = 1$ corresponds to I-VI, which does not require an initial stabilizing policy (Vrabie, 2009). With this respect, the I-GPI (Algorithm 1) contains I-PI and I-VI as two extreme cases.

Remark 1 At each i-th policy evaluation step, Algorithm 2 exactly evaluates P_{K_i} in the value function $V_{u_i}(x)$, no matter what the sampling period T_P is. That is, all the I-PIs with different sampling periods $T_p \in \mathbb{R}_+$ generate the same sequence $\{P_{K_i}\}_{i=0}^{\infty}$ and thus, can be considered the same. On the contrary, the I-GPI algorithms were considered the same only when they had the same time- and iteration-horizons T_s and k (Vrabie & Lewis, 2009). This equivalence of I-GPIs can be extended with respect to the update horizon $\hbar = kT_s$. To see this, note that according to Theorem 1, the mapping $V_{i+1}(x_t) = (\mathscr{T}_{K_i}^{T_s})^k V_i(x_t)$ in approximate policy evaluation of Algorithm 1 is equivalent to

$$V_{i+1}(x_t) = \mathscr{T}_{K_i}^{\hbar} V_i(x_t). \tag{13}$$

This corresponds to the approximation of the I-TD formula (10) and implies that *the I-GPI algorithms with the different $k \in \mathbb{N}$ but the same update horizon $\hbar = kT_s$ are all equivalent* and, hence, have the same convergence speed in the iteration domain $i \in \mathbb{Z}_+$ if it converges. Therefore, the computational complexity due to the large iteration horizon k can be lessened by increasing the time horizon T_s for the same convergence speed.

4 Monotone convergence properties of policy evaluation recursion

Vrabie & Lewis (2009) mentioned that the approximate policy evaluation of Algorithm 1 is a fixed point iteration, and proved the convergence of $V_{i|k}$ to the exact value function V_{u_i} as $k \to \infty$, under a stabilizing policy $u_i = -K_i x$. This convergence result induced the equivalence of I-PI (Algorithm 2) and I-GPI (Algorithm 1) in the limit $k \to \infty$, under an initial stabilizing policy.

We now extend this convergence property and prove the monotone convergence of $V_{i|k}$ to V_{u_i} as *the update horizon* $\hbar \in \mathbb{R}_+$ goes to infinity. Its proof is based on the following lemma which shows the several matrix iterative formulas equivalent to I-GPI. For notational convenience, we let A_i be the matrix of the i-th closed-loop system, *i.e.*, $A_i := A_{K_i}$.

Lemma 3 *Any matrices $P_{i|l}$ and $P_{i|l+\kappa}$ $(0 \le l \le l+\kappa < \infty)$ generated by i-th approximate policy evaluation of I-GPI (Algorithm 1) satisfy the following matrix equations:*

- $$P_{i|l+\kappa} = e^{A_i^T \Delta h} P_{i|l} e^{A_i \Delta h} + \int_0^{\Delta h} e^{A_i^T \tau} Q_{K_i} e^{A_i \tau}\, d\tau, \tag{14}$$
- $$P_{i|l+\kappa} - P_{i|l} = \int_0^{\Delta h} e^{A_i^T \tau} \mathscr{L}(K_i, P_{i|l}) e^{A_i \tau}\, d\tau, \tag{15}$$
- $$\mathscr{L}(K_i, P_{i|l+\kappa}) = e^{A_i^T \Delta h} \mathscr{L}(K_i, P_{i|l}) e^{A_i \Delta h}, \tag{16}$$

where $\Delta h := \kappa T_s$ and K_i is a given policy at i-th iteration, not necessarily stabilizing.

Proof. First, note that $V_{i|l+\kappa}(x_t) = \mathscr{T}_{K_i}^{\Delta h} V_{i|l}(x_t)$ holds by (12) and Theorem 1. Then, the following expansion of (9)

$$\begin{aligned}\mathscr{T}_{K_i}^{\Delta h} V_{i|l}(x_t) &= \int_0^{\Delta h} x_{t+\tau}^T Q_{K_i} x_{t+\tau}\, d\tau + x_{t+\Delta h}^T P_{i|l} x_{t+\Delta h} \\ &= x_t^T \left[\int_0^{\Delta h} e^{A_i^T \tau} Q_{K_i} e^{A_i \tau}\, d\tau + e^{A_i^T \Delta h} P_{i|l} e^{A_i \Delta h} \right] x_t,\end{aligned}$$

and the substitution of this into $V_{i|l+\kappa}(x_t) = \mathscr{T}_{K_i}^{\Delta h} V_{i|l}(x_t)$ directly proves (14). Next, adding and subtracting $P_{i|l}$ on the right hand side of (14) and using the fact that

$$e^{A_i^T \Delta h} Y e^{A_i \Delta h} - Y = \int_0^{\Delta h} e^{A_i^T \tau} (A_i^T Y + Y A_i) e^{A_i \tau}\, d\tau \tag{17}$$

holds for any matrix $Y \in \mathbb{M}^{n\times n}$, we have (15). For the proof of (16), note that (4) in Lemma 1 implies

$$\begin{aligned}&\mathscr{L}(K_i, P_{i|l+\kappa}) \\ &= \mathscr{L}(K_i, P_{i|l}) + A_i^T (P_{i|l+\kappa} - P_{i|l}) + (P_{i|l+\kappa} - P_{i|l}) A_i.\end{aligned}$$

Here, substituting (15) and using (17) with $Y = \mathscr{L}(K_i, P_{i|l})$, one can see that the second term $A_i^T (P_{i|l+\kappa} - P_{i|l}) + (P_{i|l+\kappa} - P_{i|l}) A_i$ of the right hand side is equal to $e^{A_i^T \Delta h} \mathscr{L}(K_i, P_{i|l}) e^{A_i \Delta h} - \mathscr{L}(K_i, P_{i|l})$, which completes the proof of (16). □

Based on Lemma 3, we prove the following theorem which states i) the positive definite property of the updated value function $V_{i|j}$ and ii) its (monotone) convergence to V_{u_i} in the limit $\hbar \to \infty$.

Theorem 2 *Consider the policy evaluation of Algorithm 1 at the i-th iteration. Then,*

$$P_{i|0} \geq 0 \implies P_{i|j} \geq 0 \ \forall j \in \{1,2,\cdots,k\}. \tag{18}$$

Moreover, if the policy $u_i = -K_i x$ is stabilizing, then,

(1) V_{i+1} converges to V_{u_i} as $\hbar = kT_s \to \infty$;

(2) for any $\hbar \in \mathbb{R}_+$ and $j \in \{1,2,\cdots,k\}$,

- *$\mathscr{L}(K_i, P_{i|0}) \leq 0$ implies $\mathscr{L}(K_i, P_{i|j}) \leq 0$ and*

$$P_{K_i} \leq P_{i|k} \leq \cdots \leq P_{i|j} \leq \cdots \leq P_{i|0}; \tag{19}$$

- *$0 \leq \mathscr{L}(K_i, P_{i|0})$ implies $0 \leq \mathscr{L}(K_i, P_{i|j})$ and*

$$P_{i|0} \leq \cdots \leq P_{i|j} \leq \cdots \leq P_{i|k} \leq P_{K_i}. \tag{20}$$

Proof. For the proof of (18), assume $0 \leq P_{i|0}$ and notice that Q_{K_i} is always positive semi-definite by definition. Then, obviously, both $e^{A_i^T \hbar} P_{i|0} e^{A_i \hbar}$ and $e^{A_i^T \tau} Q_{K_i} e^{A_i \tau}$ are also positive semi-definite for all h, $\tau \in \mathbb{R}_+$. Hence, (14) in Lemma 3 with $l = 0$ and $\kappa = j$ implies $P_{i|j} \geq 0$ for all $j \in \{1,2,\cdots,k\}$, which proves (18).

Next, assume that K_i is stabilizing, which implies A_i is Hurwitz. Then, the convergence of $P_{i+1} = P_{i|k}$ to P_{K_i} can be proven by showing $\mathscr{L}(K_i, P_{i|k}) \to 0$ since $P_{K_i} \geq 0$ satisfying the Lyapunov equation $\mathscr{L}(K_i, P_{K_i}) = 0$ is uniquely determined (Lancaster & Rodman, 1995, Theorem 8.5.1). Here, (16) in Lemma 3 with $l = 0$ and $\kappa = k$ (implying $\Delta h = \hbar$) guarantees the convergence $\mathscr{L}(K_i, P_{i|k}) \to 0$ since $e^{A_i \hbar}$, and thereby, $e^{A_i^T \hbar} \mathscr{L}(K_i, P_{i|0}) e^{A_i \hbar}$ converge to zero under Hurwitz A_i as $\hbar$ goes to ∞.

For the proof of monotonicity (19), suppose $\mathscr{L}(K_i, P_{i|0}) \leq 0$. Then, $e^{A_i^T t} \mathscr{L}(K_i, P_{i|0}) e^{A_i t} \leq 0$ holds for all $t \geq 0$. Therefore, we have $P_{i|1} - P_{i|0} \leq 0$ and $\mathscr{L}(K_i, P_{i|1}) \leq 0$ by (15) and (16) in Lemma 3, respectively, with $l = 0$ and $\kappa = 1$. From $\mathscr{L}(K_i, P_{i|1}) \leq 0$, we also obtain $P_{i|1} - P_{i|0} \leq 0$ in the same manner with $l = 1$ and $\kappa = 1$. Continuing this procedure up to $l = k-1$, all with $\kappa = 1$, yields (19), where the inequality $P_{K_i} \leq P_{i|k}$ is obtained from $\mathscr{L}(K_i, P_{i|k}) \leq 0$ and (15) in Lemma 3 with $l = k$ and $\kappa \to \infty$; in this limit, $P_{l+\kappa}$ converges to P_{K_i} if K_i is stabilizing. This completes the proof of (19) and the monotonicity (20) can be also proven by assuming $\mathscr{L}(K_i, P_{i|0}) \geq 0$ and following the same procedure. □

Remark 2 By Theorem 2, one can see that under stabilizing K_i, $V_{i+1}(x_t)$ obtained by approximate policy evaluation of I-GPI with finite k and T_s is an approximation of $V_{u_i}(x_t)$, and the error $|V_{i+1}(x_t) - V_{u_i}(x_t)|$ can be made arbitrarily small by increasing $\hbar$. For the limit case $\hbar \to \infty$, I-GPI under an initial stabilizing policy becomes I-PI (Algorithm 2) which generates stabilizing policies and evaluates the exact value function $V_{u_i}(x_t) = V_{i+1}(x_t)$ satisfying (7). Note that the update horizon $\hbar$ can be enlarged by either increasing the iteration horizon k or the time horizon T_s. However, the larger k is, the higher is the computational complexity; the larger T_s is, the slower the learning speed in the time domain is. Therefore, there exists a trade-off between the computational complexity and learning speed in policy evaluation, and one should carefully determine these parameters k, T_s, and of course, $\hbar$ $(= kT_s)$.

5 Stability, monotone convergence, and a new classification of I-GPI

Based on the results in Section 4, this section provides the monotone convergence and stability results of I-GPI and then establishes a new classification of I-GPI algorithms in terms of the update horizon $\hbar$. First, for notational convenience, define the increments ΔP_i, ΔK_i, and ΔK_i^* as $\Delta P_i := P_{i+1} - P_i$, $\Delta K_i := K_{i+1} - K_i$, and $\Delta K_i^* := K^* - K_{P_i}$ [2], respectively. Also, let $M_{(i,\hbar)}$ be defined by

$$M_{(i,\hbar)} := \int_0^{\hbar} e^{A_i^T \tau} \mathscr{L}(K_i, P_i) e^{A_i \tau} \, d\tau. \tag{21}$$

Then, by applying (15) and (16) in Lemma 3 to $V_{i+1}(x_t) = (\mathscr{T}_{K_i}^{T_s})^k V_i(x_t)$, with $l = 0$ and $\kappa = k$, we obtain the following two equivalent matrix formulas

$$\Delta P_i = M_{(i,\hbar)}, \tag{22}$$

$$\mathscr{L}(K_i, P_{i+1}) = e^{A_i^T \hbar} \mathscr{L}(K_i, P_i) e^{A_i \hbar}, \tag{23}$$

where $\mathscr{L}(K_i, P_i)$ for $i \geq 1$ in (23) satisfies $\mathscr{R}(P_i) = \mathscr{L}(K_i, P_i)$ due to the policy improvement step $K_i = R^{-1} B^T P_i$ of I-GPI. In addition, (5) in Lemma 1 implies that the operators $\mathscr{R}(P_{i+1})$ and $\mathscr{L}(K_i, P_{i+1})$ satisfy

$$\mathscr{R}(P_{i+1}) = \mathscr{L}(K_i, P_{i+1}) - \Delta K_i^T R \Delta K_i. \tag{24}$$

This explains how the policy improvement step $K_{i+1} = R^{-1} B^T P_{i+1}$ influences the Riccati error $\mathscr{R}(P_{i+1})$ through $\Delta K_i^T R \Delta K_i$, wherein $\mathscr{L}(K_i, P_{i+1})$ results from the policy evaluation of I-GPI and satisfies (23).

Together with all of the equivalent formulas (22)–(24) above and Lemma 2, the next lemma is essentially needed for the convergence analysis. The lemma states an additional equivalent matrix formula of I-GPI in relation to the optimal solution (K^*, P_{K^*}).

[2] K_{P_i} differs from K_i only when $i = 0$ since K_0 is arbitrarily given (see the definition of K_P and note that $K_i = R^{-1} B^T P_i$ for $i \geq 1$).

Lemma 4 *Under Assumption 1, P_i generated by I-GPI (Algorithm 1) satisfies*

$$P_{K^*} = P_i + \int_0^\infty e^{A_{K^*}^T \tau} \left[\mathscr{R}(P_i) + (\Delta K_i^*)^T R \Delta K_i^* \right] e^{A_{K^*} \tau} d\tau. \tag{25}$$

Proof. Substituting $P = P_i$, $\Phi = P_{K^*}$, and $K = K^*$ into (4) and (5) in Lemma 1, we have

$$\mathscr{L}(K^*, P_i) = A_{K^*}^T (P_i - P_{K^*}) + (P_i - P_{K^*}) A_{K^*} \tag{26}$$
$$\mathscr{L}(K^*, P_i) = \mathscr{R}(P_i) + (K^* - K_{P_i})^T R (K^* - K_{P_i}) \tag{27}$$

where $\mathscr{L}(K^*, P_{K^*}) = \mathscr{R}(P_{K^*}) = 0$ is used and the existence and uniqueness of K^* and P_{K^*} are guaranteed by Assumption 1. Substituting (27) into (26) yields the following generalized Lyapunov equation:

$$A_{K^*}^T (P_i - P_{K^*}) + (P_i - P_{K^*}) A_{K^*} = \mathscr{R}(P_i) + (\Delta K_i^*)^T R \Delta K_i^*$$

Since Assumption 1 implies that K^* is stabilizing, the proof is completed by applying Lyapunov theorem (Lancaster & Rodman, 1995, Theorem 8.5.1) to this generalized Lyapunov equation. □

In the analyses of I-GPI, it is assumed that all the P_i's generated by I-GPI are positive semi-definite. This is guaranteed under the assumption that

Assumption 2 *$P_0 \in \mathbb{M}^{n \times n}$ is positive semi-definite,*

as shown in the next proposition whose proof is trivial by (18) in Theorem 2 and mathematical induction. So, in this section, we analyze I-GPI algorithms only with Assumption 2, instead of assuming all P_i's are positive semi-definite.

Proposition 1 (Positive definiteness) *Assume that the initial matrix $P_0 \in \mathbb{M}^{n \times n}$ is positive semi-definite (resp. positive definite). Then, $P_i \in \mathbb{M}^{n \times n}$ is positive semi-definite (resp. positive definite) for all $i \in \mathbb{N}$.*

5.1 Matrix inequality conditions for stable learning

Now, we are ready to state our main theorems about I-GPI algorithms. In this subsection, several matrix inequality conditions are given, all of which are sufficient to guarantee that the updated policies are stabilizing.

Lemma 5 *Assume that K_i is stabilizing and $P_{i+1} \in \mathbb{M}^{n \times n}$ is positive semi-definite. If $\mathscr{L}(K_i, P_{i+1}) \leq Q_{K_{i+1}}$ holds for K_{i+1} updated by the policy improvement step of I-GPI (line 10 of Algorithm 1), then K_{i+1} is also stabilizing.*

Proof. See Appendix. □

Note that Assumption 2 implies that P_{i+1} is also positive semi-definite for all $i \in \mathbb{Z}_+$ by Proposition 1. So, from Lemma 5 and mathematical induction, we obtain the following stability theorem which states the general matrix bound on $\mathscr{L}(K_i, P_{i+1})$ at each iteration for stability of I-GPI.

Theorem 3 (Stability) *Suppose that K_0 is stabilizing. Then, under Assumption 2, if K_i and P_{i+1} satisfy $\mathscr{L}(K_i, P_{i+1}) \leq Q_{K_{i+1}}$ for all $i \in \mathbb{Z}_+$, then, K_i is stabilizing for all $i \in \mathbb{Z}_+$.*

Proof. Assume K_i is stabilizing and satisfies $\mathscr{L}(K_i, P_{i+1}) \leq Q_{K_{i+1}}$. Then, the application of Lemma 5 with $P_{i+1} \geq 0$ implies that K_{i+1} is also stabilizing and mathematical induction concludes that K_i is stabilizing for all $i \in \mathbb{Z}_+$. □

The condition $\mathscr{L}(K_i, P_{i+1}) \leq Q_{K_{i+1}}$ in Theorem 3 provides stability during and after the learning phase, but the direct evaluation of $\mathscr{L}(K_i, P_{i+1})$ requires the knowledge of the matrix A at each iteration, while I-GPI does not. The next theorem provides another inequality condition for the closed-loop stability, which does not explicitly depend on $\mathscr{L}(K_i, P_{i+1})$.

Theorem 4 (Stability) *Suppose that Assumption 2 holds and K_0 is a stabilizing policy that satisfies $\mathscr{L}(K_0, P_0) \leq Q_{K_0}$. Then, K_i is stabilizing for all $i \in \mathbb{Z}_+$ if K_i and K_{i+1} satisfy*

$$e^{A_i^T \hbar} Q_{K_i} e^{A_i \hbar} \leq Q_{K_{i+1}}, \quad \forall i \in \mathbb{Z}_+. \tag{28}$$

Proof. Assume that K_i is stabilizing and satisfies $\mathscr{L}(K_i, P_i) \leq Q_{K_i}$. Then, we have from (23) and (28)

$$\mathscr{L}(K_i, P_{i+1}) = e^{A_i^T \hbar} \mathscr{L}(K_i, P_i) e^{A_i \hbar} \leq e^{A_i^T \hbar} Q_{K_i} e^{A_i \hbar} \leq Q_{K_{i+1}}.$$

So, since $P_{i+1} \geq 0$ holds (see Assumption 2 and Proposition 1), K_{i+1} is also stabilizing by Lemma 5. Substituting the inequality into (24) and rearranging it yield

$$\mathscr{L}(K_{i+1}, P_{i+1}) = \mathscr{R}(P_{i+1}) \leq Q_{K_{i+1}} - \Delta K_i^T R \Delta K_i \leq Q_{K_{i+1}}.$$

Therefore, mathematical induction with $\mathscr{L}(K_0, P_0) \leq Q_{K_0}$, proves that K_i is stabilizing for all $i \in \mathbb{Z}_+$. □

Remark 3 Although condition (28) depends on the system matrix A, it is contained only in the form of exponentials. By virtue of this fact, (28) can be easily checked without knowing the system matrix A. This issue will be further discussed in detail in Section 6, where a data-driven method is proposed to check the condition (28) without explicit use of the knowledge of A.

5.2 PI-mode of convergence and stability

In Section 3, we mentioned that in the limit $\hbar \to \infty$, I-GPI under an initial stabilizing policy is equivalent to I-PI (Algorithm 2) which generates the sequence $\{K_i\}_{i\in\mathbb{N}}$ of stabilizing policies and their exact value function matrices $\{P_{K_i}\}_{i\in\mathbb{Z}_+}$. Moreover, the equivalence to Kleinman (1968)'s Newton method (Vrabie, 2009) implies that I-PI guarantees 2nd-order monotone convergence of P_{K_i} to P_{K^*} with decreasing order

$$0 \le P_{K^*} \le \cdots \le P_{K_{i+1}} \le P_{K_i} \le \cdots \le P_{K_0}.$$

In the following, we state PI-mode of convergence of I-GPI, which implies that, under certain conditions, K_i generated by I-GPI is stabilizing for all i, and P_i monotonically converges to P_{K^*} in a similar manner to I-PI.

Theorem 5 (PI-mode of convergence) *Suppose the initial policy K_0 and the initial matrix P_0 satisfy $\mathscr{L}(K_0,P_0) \le 0$. Then, under Assumptions 1–2, the following hold for all $i \in \mathbb{Z}_+$:*

- ***(Stability)*** *K_i is stabilizing and satisfies the Lyapunov inequalities $\mathscr{L}(K_i,P_{i+1}) \le 0$ and $\mathscr{L}(K_{i+1},P_{i+1}) \le -\Delta K_i^T R \Delta K_i$. That is,*

$$\begin{cases} A_i^T P_{i+1} + P_{i+1} A_i \le -Q_{K_i} \\ A_{i+1}^T P_{i+1} + P_{i+1} A_{i+1} \le -Q_{K_{i+1}} - \Delta K_i^T R \Delta K_i. \end{cases} \tag{29}$$

- ***(Monotone convergence from above)*** *The sequences $\{P_i\}_{i=0}^\infty$ and $\{K_i\}_{i=0}^\infty$ converge to the optimal solution P_{K^*} and K^*, respectively, with the following monotonicities:*

$$\begin{cases} 0 \le P_{K_i} \le P_{i+1} \le P_i \\ 0 \le P_{K^*} \le \cdots \le P_{i+1} \le P_i \le \cdots \le P_0. \end{cases} \tag{30}$$

- ***(2^{nd}-order monotone decreasing)*** *There exists $c > 0$ such that if (P_{i+1}, K_{i+1}) and K_i for some $i \in \mathbb{Z}_+$ satisfy*

$$-(\alpha_i - 1)\Delta K_i^T R \Delta K_i \le \mathscr{L}(K_i, P_{i+1}) \le 0, \tag{31}$$

where $\alpha_i \ge 1$ is a constant, then, for all such i,

$$\|P_{i+1} - P_{K^*}\| \le c \cdot \alpha_i \|P_i - P_{K^*}\|^2. \tag{32}$$

Proof. First, assume $\mathscr{L}(K_i,P_i) \le 0$ for some $i \in \mathbb{Z}_+$. Then, we have $\mathscr{L}(K_i,P_{i+1}) \le 0$ by (23) and substituting this into (24) yields

$$\mathscr{R}(P_{i+1}) = \mathscr{L}(K_i,P_{i+1}) - \Delta K_i^T R \Delta K_i \le -\Delta K_i^T R \Delta K_i.$$

Therefore, mathematical induction with (K_0,P_0) satisfying $\mathscr{L}(K_0,P_0) \le 0$ implies the Lyapunov inequalities $\mathscr{L}(K_i,P_{i+1}) \le 0$ and $\mathscr{L}(K_{i+1},P_{i+1}) = \mathscr{R}(P_{i+1}) \le -\Delta K_i^T R \Delta K_i \le 0$ hold for all $i \in \mathbb{Z}_+$.

(Proof of stability). $\mathscr{L}(K_i,P_{i+1}) \le 0$ implies $\mathscr{L}(K_i,P_{i+1}) \le Q_{K_{i+1}}$, and by Theorem 3 with Assumption 2, one can conclude that for all $i \in \mathbb{Z}_+$, K_i is stabilizing.

(Proof of monotone convergence). Proposition 1 with Assumption 2 guarantees $P_i \ge 0\ \forall i \in \mathbb{Z}_+$. So, (19) in Theorem 2 implies that P_{i+1} satisfies $0 \le P_{K_i} \le P_{i+1} \le P_i$, which holds for all $i \in \mathbb{Z}_+$ ($\because \mathscr{L}(K_i,P_i) \le 0$ *for all* $i \in \mathbb{Z}_+$). Therefore, since it is monotonically decreasing and bounded by 0, the sequence $\{P_i\}_{i=0}^\infty$ monotonically converges, and thereby, $\{K_i\}_{i=0}^\infty$ also converges ($\because K_i = R^{-1}B^T P_i\ \forall i \in \mathbb{N}$). Let $\bar{P}$ and $\bar{K}$ be their respective limit points, *i.e.*, $\bar{P} = \lim_{i\to\infty} P_i$ and $\bar{K} = \lim_{i\to\infty} K_i$. Then, $\bar{P}$ satisfies

$$0 \le \bar{P} \le P_{i+1} \le P_i \quad \text{for all } i \in \mathbb{Z}_+,$$

and Lemma 2 implies $\bar{P} = P_{K^*}$ and $\bar{K} = K^*$, respectively, which proves the convergence with the monotonicity (30).

(Proof of 2^{nd}-order convergence). First, note that one has $0 \le P_{i+1} - P_{K^*}$ by the monotone convergence (30), and the equations (24) and (31) imply

$$-\alpha_i \Delta K_i^T R \Delta K_i \le \mathscr{R}(P_{i+1}) \le -\Delta K_i^T R \Delta K_i.$$

From this inequality and (25), one obtains

$$\begin{aligned} 0 \le P_{i+1} - P_{K^*} &\le -\int_0^\infty e^{A_{K^*}^T \tau} \mathscr{R}(P_{i+1}) e^{A_{K^*}\tau} d\tau \\ &\le \alpha_i \int_0^\infty e^{A_{K^*}^T \tau} \Delta K_i^T R \Delta K_i e^{A_{K^*}\tau} d\tau. \end{aligned} \tag{33}$$

By virtue of the fact that

$$0 \le X \le Y \text{ for } X, Y \in \mathbb{M}^{n\times n} \implies \|X\| \le \|Y\|,$$

one can take the matrix norm $\|\cdot\|$ on (33) and obtain the following inequality using the properties of the norm:

$$\begin{aligned} \|P_{i+1} - P_{K^*}\| &\le \alpha_i \int_0^\infty \left\| e^{A_{K^*}^T \tau} \Delta K_i^T R \Delta K_i e^{A_{K^*}\tau} \right\| d\tau \\ &\le \alpha_i \underbrace{\left(\int_0^\infty \|e^{A_{K^*}\tau}\|^2 d\tau \right) \|BR^{-1}B^T\|}_{=:c} \cdot \|P_i - P_{i+1}\|^2 \\ &= c \cdot \alpha_i \|P_i - P_{i+1}\|^2. \end{aligned}$$

Now, the proof of the 2^{nd}-order monotone decreasing (32) can be done by using the fact that by the monotonicity (30), $0 \le P_i - P_{i+1} \le P_i - P_{K^*}$ holds for all $i \in \mathbb{Z}_+$, which again implies $\|P_i - P_{i+1}\| \le \|P_i - P_{K^*}\|$. □

Remark 4 The properties of I-GPI in PI-mode of convergence shown in Theorem 5 are similar to those of I-PI which is equivalent to Kleinman (1968)'s Newton method. Actually, the I-GPI algorithm can be considered an inexact Kleinman's Newton algorithm (Feitzinger *et al.*, 2009) with the

residual "$\mathscr{L}(K_i,P_{i+1})$", which can be made arbitrarily small by increasing $\hbar$ in the policy evaluation step. Moreover, denoting the Frechet derivative of $\mathscr{L}(K_i,P_i)$ taken with respect to P_i by $\mathscr{L}'_{K_i,P_i}$, we can see that I-GPI is also equivalent to the quasi-Newton method updated by

$$P_{i+1} = P_i + \left(\mathscr{L}'_{K_i,P_i}\right)^{-1}\left[\mathscr{L}(K_i,P_i) - \mathscr{L}(K_i,P_{i+1})\right],$$

which converges to the Newton method as the residual $\mathscr{L}(K_i,P_{i+1})$ goes to zero ($\hbar\to\infty$). In this limit case, I-GPI becomes I-PI as mentioned in Section 4, and the Lyapunov inequalities in (29) of Theorem 5 become their respective Lyapunov equations

$$\begin{cases} A_i^T P_{i+1} + P_{i+1}A_i = -Q_{K_i} \\ A_{i+1}^T P_{i+1} + P_{i+1}A_{i+1} = -Q_{K_{i+1}} - \Delta K_i^T R \Delta K_i, \end{cases} \tag{34}$$

which provide the closed-loop stability for all $i\in\mathbb{Z}_+$. Moreover, (34) implies that i) the residual $\mathscr{L}(K_i,P_{i+1})$ becomes zero and ii) $\mathscr{R}(P_{i+1})$ satisfies $\mathscr{R}(P_{i+1}) = -\Delta K_i R \Delta K_i$ for all $i\in\mathbb{Z}_+$. Here, the former guarantees the condition (31) with $\alpha_i = 1$, which implies the uniform 2^{nd}-order monotone convergence of I-PI, and the latter implies $\mathscr{R}(P_i)\le 0$, which provides an alternative approach to the proof of monotone convergence of I-PI, as shown in this paper.

5.3 Local monotone convergence under stable closed-loop dynamics

Next, we investigate local monotone decreasing and convergence of I-GPI near the solution P_{K^*} satisfying $\mathscr{R}(P_{K^*}) = 0$. For notational convenience, let $\Phi_{(i,\hbar)}$ be defined as

$$\Phi_{(i,\hbar)} := \int_0^\hbar \|e^{A_i\tau}\|^2\, d\tau. \tag{35}$$

In the following, the convergence property will be discussed only in the local set Ω_i^r $(r=1,2)$ defined by

$$\Omega_i^r := \left\{P_i\in\mathbb{M}^{n\times n} : \rho_{(r,i)} < 1 - \|e^{A_i\hbar}\|^2/\|\mathscr{L}(K_i,P_i)\|^{r-1}\right\}$$

where $\rho_{(r,i)} := \|\Delta K_i^T R \Delta K_i\|/\|\mathscr{L}(K_i,P_i)\|^r$. Here, it can be easily derived from the definitions of Ω_i^r and $\rho_{(r,i)}$ $(r=1,2)$ that

$$\begin{cases} \|\mathscr{L}(K_i,P_i)\| < 1 \Longrightarrow \Omega_i^2 \subset \Omega_i^1 \\ \|\mathscr{L}(K_i,P_i)\| \ge 1 \Longrightarrow \Omega_i^1 \subseteq \Omega_i^2. \end{cases}$$

Moreover, for having both $\Omega_i^1\neq\phi$ and $\Omega_i^2\neq\phi$, $\|e^{A_i\hbar}\|<1$ should be satisfied, which is attainable with a sufficiently large $\hbar\in\mathbb{R}_+$ when A_i is Hurwitz.

Theorem 6 (Local monotone convergence) *Assume that A_i is Hurwitz for all $i\in\mathbb{Z}_+$. Then, under Assumption 1,*

- ***(Monotone decreasing)** if $P_i\in\Omega_i^r$ for $r\in\{1,2\}$, then, $\mathscr{R}(P_{i+1})$ satisfies*

$$\|\mathscr{R}(P_{i+1})\| \le \|\mathscr{L}(K_i,P_i)\|^r. \tag{36}$$

- ***(Monotone convergence)** if $P_i\in\Omega_i^1$ for all $i\in\mathbb{Z}_+$, then, $\{P_i\}_{i=0}^\infty$ converges to P_{K^*} with the 1^{st}-order monotonicity (36) $(r=1)$.*

Proof. Taking the matrix norm $\|\cdot\|$ of (24), using the basic properties of the norm with the substitution of (23), and rearranging the equation yield the following for $r=1,2$:

$$\begin{aligned} \|\mathscr{R}(P_{i+1})\| &\le \|e^{A_i\hbar}\|^2\|\mathscr{L}(K_i,P_i)\| + \|\Delta K_i R \Delta K_i\| \\ &= \left[\frac{\|e^{A_i\hbar}\|^2}{\|\mathscr{L}(K_i,P_i)\|^{r-1}} + \rho_{(r,i)}\right]\|\mathscr{L}(K_i,P_i)\|^r. \end{aligned} \tag{37}$$

For $\|\mathscr{R}(P_{i+1})\| < \|\mathscr{L}(K_i,P_i)\|^r$, the value in brackets in (37) should be less than 1, which corresponds to $P_i\in\Omega_i^r$, the completion of the proof of monotone decreasing.

Next, assume that $P_i\in\Omega_i^1$ for all $i\in\mathbb{Z}_+$. Then, by the above argument, $\|\mathscr{R}(P_i)\|$ is monotonically decreasing and bounded by '0' as follows:

$$0\le\cdots\le\|\mathscr{R}(P_{i+1})\|\le\|\mathscr{R}(P_i)\|\le\cdots\le\|\mathscr{L}(K_0,P_0)\|$$

where $K_i = R^{-1}B^TP_i$ is substituted for all $i\in\mathbb{N}$. This implies $\{\mathscr{R}(P_i)\}$ and hence $\{P_i\}$ are convergence sequences under this norm, and so is $\{K_i\}$ as well. Therefore, denoting $\bar{P} = \lim_{i\to\infty}P_i$ and $\bar{K} = \lim_{i\to\infty}K_i$, we can conclude $\bar{P} = P_{K^*}$ and $\bar{K} = K^*$ by Lemma 2, which completes the proof of the local monotone convergence. □

Remark 5 To enlarge the convergence region Ω_i^r $(r=1,2)$, both $\|e^{A_i\hbar}\|^2 \ll 1$ and $\rho_{(r,i)}\approx 0$ are necessary. The former can be achieved by policy evaluation with a Hurwitz matrix A_i and a sufficiently large update horizon $\hbar$. Especially, we obtain $\|e^{A_i\hbar}\|^2\to 0$ in the case of I-PI ($\hbar\to\infty$). On the other hand, the latter $\rho_{(r,i)}\approx 0$ can be obtained when the policy K_i is almost stationary (*i.e.* $\Delta K_i\approx 0$). Note that, if $\Delta K_i = 0$, policy improvement does not contribute to the variations of P_i, and hence, I-GPI becomes equal to its policy evaluation step (no policy improvement). Therefore, in the case of $\Delta K_i = 0$, we have $\mathscr{R}(P_{i+1}) = \mathscr{L}(K_i,P_{i+1})$, and the global monotone convergence is achieved by Theorem 2.

Remark 6 From the definitions of $M_{(i,\hbar)}$ and $\Phi_{(i,\hbar)}$ (see (21) and (35)), one can easily derive the inequality $\|M_{(i,\hbar)}\| \le \Phi_{(i,\hbar)}\|\mathscr{L}(K_i,P_i)\|$. This again yields

$$\rho_{(r,i)} \le \Phi_{(i,\hbar)}^2\|B^TR^{-1}B\|\|\mathscr{L}(K_i,P_i)\|^{2-r}.$$

where $\Delta K_i^T R\Delta K_i = \Delta P_i B R^{-1} B^T \Delta P_i$ and (22) are used (see also (21)). By substituting the right hand side of this into

$\rho_{(r,i)}$ in Ω_i^r and rearranging the result, the local regions $\underline{\Omega}_i^r \subseteq \Omega_i^r$ $(r=1,2)$ for r-th order decreasing/convergence can be obtained as follows:

$$\underline{\Omega}_i^1 := \{P_i \in \mathbb{M}^{n\times n} : \|\mathscr{L}(K_i,P_i)\| < D_i\},$$
$$\underline{\Omega}_i^2 := \{P_i \in \mathbb{M}^{n\times n} : 0 \le E_i < \|\mathscr{L}(K_i,P_i)\| < 1\}, \quad (38)$$

where $D_i := \dfrac{1-\|e^{A_i\hbar}\|^2}{\|BR^{-1}B^T\|\Phi^2_{(i,\hbar)}}$, $E_i := \dfrac{\|e^{A_i\hbar}\|^2}{1-\|BR^{-1}B^T\|\Phi^2_{(i,k)}}$.

Here, it can be easily checked that

$$1 < D_i \text{ and } \|BR^{-1}B^T\|\Phi^2_{(i,\hbar)} < 1$$

imply $0 \le E_i < 1$, so guarantee the existence of the nonempty set $\underline{\Omega}_i^2 \neq \phi$ (just rearrange the inequality $1 < D_i$). Although these sets $\underline{\Omega}_i^1$ and $\underline{\Omega}_i^2$ are rather conservative, they provide the *concrete local ranges* for the local monotone convergence. Especially, (38) guarantees 2nd-order local decreasing up to E_i under $\|BR^{-1}B^T\|\Phi^2_{(i,\hbar)} < 1$ *without the assumption of* $\mathscr{L}(K_0,P_0) \le 0$. For the I-PI case ($\hbar \to \infty$), E_i converges to zero since $\lim_{\hbar\to\infty}\|e^{A_i\hbar}\| = 0$ ($\because A_i$ is Hurwitz in I-PI). More detailed discussions regarding $\underline{\Omega}_i^1$ and $\underline{\Omega}_i^2$ are shown in our preliminary result (Lee *et al.*, 2011).

5.4 Monotone increasing and VI-mode of convergence

It has been shown that VI for DT LQR has the following monotone increasing convergence property (Lendelius, 1997; Lewis & Vrabie, 2009; Zhang *et al.*, 2009):

$$\begin{cases} \lim_{i\to\infty} P_i = P_{K^*} \text{ and} \\ 0 = P_0 \le P_1 \le \cdots \le P_i \le P_{i+1} \le \cdots \le P_{K^*} \end{cases} \quad (39)$$

under zero initial value function $V_0(x) \equiv 0$ ($P_0 = 0$), where $\{P_i\}_{i\in\mathbb{Z}_+}$ is generated by VI for DT LQR. In this case, the initial policy is not necessarily stabilizing.

We call this kind of monotone convergence "VI-mode of convergence", the counter part of PI-mode of convergence (see Fig. 2(a)). In the following, we discuss monotone increasing and this VI-mode of convergence of I-GPI methods.

Theorem 7 *Under Assumptions 1–2 with $0 \le P_0 \le P_{K^*}$, if K_i and P_i satisfy $\mathscr{L}(K_i,P_i) \ge 0$, $\forall i \in \{0,1,2,\cdots,l-1\}$, then $\{P_i\}_{i=0}^l$ possesses the monotone increasing property:*

$$0 \le P_0 \le \cdots \le P_i \le P_{i+1} \le \cdots \le P_l \le P_{K^*}. \quad (40)$$

Moreover, if $\mathscr{L}(K_i,P_i) \ge 0$ for all $i \in \mathbb{Z}_+$, then, $\{P_i\}_{i=0}^\infty$ converges to P_{K^*} with the monotonicity (40) $\forall i \in \mathbb{Z}_+$.

Proof. For each $i \in \{0,1,2,3,\cdots,l-1\}$, $\mathscr{L}(K_i,P_i) \ge 0$ and (20) in Theorem 2 implies $0 \le P_i \le P_{i+1}$ (positive semi-definiteness comes from Proposition 1). Next, we obtain $P_{K^*} \ge P_i$ $\forall i \in \{1,2,\cdots,l-1\}$ from (25) since $\mathscr{R}(P_i) = \mathscr{L}(K_i,P_i) \ge 0$ by assumption and $(\Delta K_i^*)^T R \Delta K_i^* \ge 0$ ($P_{K^*} \ge P_0$ is assumed for $i=0$). Rearranging all these inequalities yields

$$0 \le P_i \le P_{i+1} \le P_{K^*},$$

which holds for all $i \in \{0,1,2,\cdots,l-1\}$ by the assumption $\mathscr{L}(K_i,P_i) \ge 0$ for all such i. Therefore, we have (40), and the monotone convergence to the optimal solution can be directly proven by the assumption of $\mathscr{L}(K_i,P_i) \ge 0$ for all $i \in \mathbb{Z}_+$ and Lemma 2. □

This theorem with $0 \le P_0 \le P_{K^*}$ and $\mathscr{L}(K_0,P_0) \ge 0$ obviously guarantee the monotone increasing (40) up to some finite $l \in \mathbb{N}$ (the trivial case is $l=1$). For I-GPI methods with $P_0 = 0$, this monotone increasing is also valid for any given initial policy K_0 not necessarily stabilizing since $\mathscr{L}(K_0,P_0) = K_0^T R K_0 + Q \ge 0$ holds. On the other hand, for VI-mode of convergence, I-GPI should generate the sequences $\{P_i\}$ and $\{K_i\}$ $(= R^{-1}B^T P_i)$, both of which satisfy $\mathscr{L}(K_i,P_i) \ge 0$ for all $i \in \mathbb{N}$. However, this is not attainable in general since, even in the case where K_i is stabilizing that satisfies $\mathscr{L}(K_i,P_i) \ge 0$ under $0 \le P_i \le P_{K^*}$, $\mathscr{R}(P_{i+1})$ can be indefinite or negative semi-definite for large $\hbar$ by (24). This is because the residual $\mathscr{L}(K_i,P_{i+1})$ in (24) becomes zero as $\hbar \to \infty$ (see (23) and also Theorem 2). More obviously and intuitively, since P_{K^*} is the optimal solution, any P_{K_i} for a stabilizing K_i satisfies $0 \le P_{K^*} \le P_{K_i}$, which in turn implies that P_i would not satisfy $0 \le P_i \le P_{K^*}$ especially when $\hbar$ is large (an example of this case is $P_{K^*} \le P_i \le P_{K_i}$). Therefore, VI-mode of convergence is not attainable in general.

In contrast, I-GPI methods with a sufficiently small $\hbar > 0$ can generate the sequence $\{P_i\}$, which satisfies $\mathscr{L}(K_i,P_i) \ge 0$ for all $i \in \mathbb{N}$ and hence, converges in VI-mode by Theorem 7. In this case, (P_0,K_0) is required to satisfy $\mathscr{L}(K_0,P_0) > 0$, instead of $\mathscr{L}(K_0,P_0) \ge 0$.. To see this, assume $\mathscr{L}(K_i,P_i) > 0$. Then, $\mathscr{L}(K_i,P_{i+1})$ is also positive definite by (23) or Theorem 2, which again implies there is $\varepsilon_i > 0$ such that $\varepsilon_i I_n \le \mathscr{L}(K_i,P_{i+1})$. So, if ΔK_i satisfies

$$\Delta K_i^T R \Delta K_i < \varepsilon_i I_n, \quad (41)$$

then P_{i+1} satisfies $\mathscr{R}(P_{i+1}) > 0$ by (24); the induction implies $\mathscr{R}(P_i) > 0$ for all $i \in \mathbb{N}$, and the VI-mode of convergence is guaranteed by Theorem 7. Here, since $\left\|\Delta K_i^T R \Delta K_i\right\|$ can be made arbitrarily small by decreasing $\hbar > 0$ (notice $\Delta K_i = R^{-1}B^T M_{(i,\hbar)}$ by (22) and $M_{(i,\hbar)} \to 0$ as $\hbar \to 0$), the I-GPI with a sufficiently small $\hbar > 0$ yields ΔK_i satisfying (41) and thereby, can generate the convergent sequence $\{P_i\}$ in VI-mode by Theorem 7.

This VI-mode of convergence can be also possible for I-VI with sufficiently small $T_s > 0$ since it belongs to the special class of I-GPI with $k=1$ and sufficiently small $\hbar > 0$.

Actually, it can be shown that the infinitesimal version of I-GPI ($\hbar \to 0$) is governed by the forward-in-time differential Riccati equation (DRE) $\dot{P}_t = \mathscr{R}(P_t)$ $(0 \le t < \infty)$ which can be obtained by dividing both sides of (22) by $\hbar$ and limiting $\hbar \to 0$ (Vrabie, 2009). In this case, under the zero initial condition $P_0 = 0$, which is the special case of $0 \le P_0 \le P_{K^*}$, and $\mathscr{R}(P_0) \ge 0$ shown in Theorem 7, Lancaster & Rodman (1995) showed that P_t generated by the forward-in-time DRE monotonically converges to P_{K^*} with the monotonicity $0 \le P_{t_1} \le P_{t_2} \le P_{K^*}$ for $t_1 \le t_2$ (Lancaster & Rodman, 1995, Theorem 16.4.3). Theorem 7 and the above discussions imply that this monotone convergence result can be extended to the general case "$0 \le P_0 \le P_{K^*}$ and $\mathscr{R}(P_0) \ge 0$."

5.5 *Summary and the new classification of I-GPI*

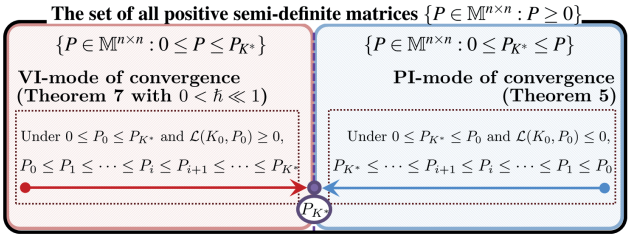
So far, we have shown five properties regarding monotone convergence and/or stability of I-GPI methods, and the two main results are PI- and VI-mode of convergence, which correspond to monotone decreasing and increasing convergence, respectively. Fig. 2(a) describes these two modes of convergence of I-GPI.

- In PI-mode, P_i remains in the region $\{0 \le P_{K^*} \le P\}$ for all $i \in \mathbb{Z}_+$ and converges like PI methods, *e.g.*, I-PI for CT LQR (I-GPI in the limit $\hbar \to \infty$).
- In VI-mode, P_i is in the other region $\{0 \le P \le P_{K^*}\}$ for all $i \in \mathbb{Z}_+$, and converges like VI for DT LQR.

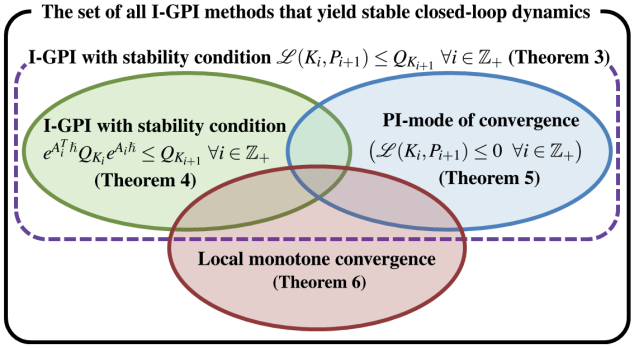
Note that PI- and VI-mode of convergence of I-GPI can be considered the two counterparts of (19) and (20) in Theorem 2 describing monotone convergence of each policy evaluation step. While the choice of the update horizon $\hbar > 0$ does not affect PI-mode of convergence (Theorem 5), VI-mode of convergence can be achieved only with sufficiently small $\hbar > 0$ (or in the limit $\hbar \to 0$) as discussed in Section 5.4; otherwise, $\mathscr{L}(K_i, P_i) \ge 0$ is not guaranteed after some finite step $i = l$ and in this case, Theorem 7 only implies I-GPI generates P_i that is monotonically increasing up to l. On the other hand, PI-mode of convergence is possible even in the limit $\hbar \to 0$ and $\hbar \to \infty$ as long as K_0 and $P_0 \ge 0$ satisfy $\mathscr{L}(K_0, P_0) \le 0$.

From the above discussions, infinitesimal GPI ($\hbar \to 0$) with $0 \le P_0 \le P_{K^*}$ and I-PI ($\hbar \to \infty$ under stabilizing K_0) can be considered the representatives of I-GPI in VI- and PI-mode of convergence, respectively. It is also shown in Section 3 that I-GPI with the same $\hbar$ are all equal in iteration domain (see Remark 1). From these facts, we establish a new classification of I-RL (and I-GPI) methods with respect to the update horizon $\hbar$, as shown in Fig. 1, where infinitesimal GPI is at one extreme tip ($\hbar \to 0$), and I-PI is at the other extreme tip of the spectrum ($\hbar \to \infty$). I-VI ($k = 1$) and I-GPI (fininte k) are posed on the middle of the spectrum, and their convergence properties are determined depending on the update horizon $\hbar$ and conditions presented in this section.

Unlike VI-mode, I-GPI in PI-mode of convergence guarantees the closed-loop stability for all $i \in \mathbb{Z}_+$ (Theorem 5).



(a) PI- and VI-mode of convergence



(b) Stability and monotone convergence

Fig. 2. Summary of stability and monotone convergence properties of I-GPI.

With the assumption that K_0 is stabilizing, three types of matrix inequalities have been presented for stability of I-GPI, namely, $\mathscr{L}(K_i, P_{i+1}) \le Q_{K_{i+1}}$ in Theorem 3, (28) in Theorem 4, and $\mathscr{L}(K_i, P_{i+1}) \le 0$ in Theorem 5 (PI-mode of convergence). As can be seen from Fig. 2(b),

- the first one is the most general condition for stability, and hence can be considered the sufficient condition of the other two;
- the second one (28) is rather restricted, but can be checked in online learning without using the knowledge of the matrix A (see the next section for this issue).
- The last one $\mathscr{L}(K_i, P_{i+1}) \le 0$ for PI-mode of convergence obviously implies $\mathscr{L}(K_i, P_{i+1}) \le Q_{K_{i+1}}$ and thereby guarantees the closed-loop stability. Moreover, K_i and P_i generated by I-GPI in PI-mode satisfy $\mathscr{L}(K_i, P_{i+1}) \le 0$, so the agent does not need to check any matrix inequality for stability, except $\mathscr{L}(K_0, P_0) \le 0$ at $i = 0$ (see Theorem 5). For the first two conditions, the agent should check the inequality at every step to guarantee the stability.

Fig. 2(b) also illustrates that local monotone convergence in Theorem 6 is achieved under the stable closed-loop dynamics, but it is not restricted or affected by any matrix inequalities in Theorems 3–5. Instead, P_i should be near the solution P_{K^*}, *i.e.*, $P_i \in \{P \in \mathbb{M}^{n\times n} : \|\mathscr{R}(P)\| \le \varepsilon\}$ for sufficiently small $\varepsilon > 0$ (see $\underline{\Omega}_1$ in Remark 6). In this case, the norm of $\mathscr{R}(P_i)$ monotonically decreases and eventually converges to '0' $(= \mathscr{R}(P_{K^*}))$ (see Theorem 6). Here, sufficiently large $\hbar > 0$ under stable closed-loop dynamics enlarges the region of convergence $\underline{\Omega}_1$ (or Ω_1) since it makes $\left\|e^{A_i \hbar}\right\| \approx 0$.

6 Data-driven Implementation

Based on the mathematical results, we propose in this section data-driven methods for implementing I-GPI (Algorithm 1), specifically, the LS methods for uniquely evaluating $P_{i|j}$ satisfying (12) and determining k at each iteration, without violating the matrix inequalities given in Section 5. The proposed method can be also used for implementing the I-GPI given by Vrabie & Lewis (2009). Let $t_i > 0$ be the time at which the i-th step of Algorithm 1 is started. Then, for $\tau > 0$ define $W_i(\tau)$ as the solution of the differential equation

$$\dot{W}_i(\tau) = x_{t_i+\tau}^T Q_{K_i} x_{t_i+\tau}, \;\; W_i(0) = 0.$$

The method is based on the data pairs $(x_i[N], W_i[N])$ sampled at i-th iteration with the same period T_s and the same i-th policy $u_i = -K_i x$, where $x_i[N]$ and $W_i[N]$ are defined as

$$x_i[N] := x_{t_i+NT_s}, \;\; W_i[N] := W_i(NT_s),$$

respectively, for $N = 0, 1, 2, \cdots, N_{i,max}$. Here, $N_{i,max}$ denotes the number of data pairs $(x_i[N], W_i[N])$ the I-GPI agent collected. For $N = 0$, since $W_i[0] = 0$ by the zero-initial condition, $W_i[0]$ does not need to be sampled, but $x_i[0]$ is additionally measured by the agent. Therefore, the numbers of the sampled data $x_i[N]$ and $W_i[N]$ at each iteration will be $N_{i,max}+1$ and $N_{i,max}$, respectively.

For notational convenience, let $x_{i,j} \in \mathbb{R}$ be the j-th element of $x_i[N]$, and $x_{i,j:n} \in \mathbb{R}^{n-j+1}$ be the column vector defined as $x_{i,j:n} := [x_{i,j} \;\; x_{i,j+1} \;\; \cdots \;\; x_{i,n}]^T$. We also denote the minimum number of the sampled data pairs $(x_i[N], W_i[N])$ required to implement Algorithm 1 by $N_{i,min}$. Then, the number of the measured data "$N_{i,max}$" at the i-th iteration should be larger than or equal to $N_{i,min}$, *i.e.*, $N_{i,min} \le N_{i,max}$ for reasonable implementation. This minimum requirement $N_{i,min}$ is actually connected to the certain excitation condition that should be satisfied for the computation of $V_{i|j+1}(x)$, $(j = 0, 1, 2, \cdots k-1)$ by (12).

Assumption 3 (Excitation condition) *For each i-th step, there exist $N_{i,min} \in \mathbb{N}$, $\underline{\alpha} \in \mathbb{R}_+$, and $\bar{\alpha} \in \mathbb{R}_+$ such that*

$$\underline{\alpha} I \le \sum_{N=0}^{N_{i,max}-1} \bar{x}_i[N] \bar{x}_i^T[N] \le \bar{\alpha} I, \tag{42}$$

for all $N_{i,max} \ge N_{i,min}$, where $\bar{x}_i[N] \in \mathbb{R}^{n(n+1)/2}$ is the Kronecker product quadratic polynomial basis vector of $x_i[N]$, that is, $\bar{x}_i[N] := \left[x_{i,1}x_{i,1:n}^T \;\; x_{i,2}x_{i,2:n}^T \;\; x_{i,3}x_{i,3:n}^T \;\; \cdots \;\; x_{i,n}^2\right]^T$.

For description of the method, let $\bar{X}_{i,p}$ $(p = 0, 1)$ be the matrix of dimension $n(n+1)/2$-by-$N_{i,max}$, which is made by column-wisely collecting $\bar{x}_i[N]$, that is,

$$\bar{X}_{i,p} := \left[\bar{x}_i[p] \;\; \bar{x}_i[p+1] \;\; \bar{x}_i[p+2] \;\; \cdots \;\; \bar{x}_i[p+N_{i,max}-1]\right].$$

Then, one can see that the excitation condition (42) can be rewritten in terms of $\bar{X}_{i,0}$ as "$\underline{\alpha} I \le \bar{X}_{i,0}\bar{X}_{i,0}^T \le \bar{\alpha} I$", which implies $\bar{X}_{i,0}$ is of full rank and $\bar{X}_{i,0}\bar{X}_{i,0}^T$ is invertible. Note that, for $\bar{X}_{i,0}$ to be full rank, $N_{i,min}$ needs to be larger than or equal to the dimension of $\bar{x}_i[N]$, '$n(n+1)/2$'. So, in the implementation method, the agent should collect at least $n(n+1)/2$ data points at each iteration for satisfying Assumption 3.

6.1 Implementation of policy evaluation with a finite k

Recall that the basic operation of I-GPI is the one-step DP recursion (12); at each policy evaluation step of Algorithm 1, the agent performs the one-step DP recursion k-times. In the following, we present a sub-algorithm which determines $P_{i|j+1}$ based on the quantities $P_{i|j}$, $x_i[0]$, and the sampled data $(x_i[N], W_i[N])$ for $N = 1, 2, \cdots, N_{i,max}$. To derive the method, note that, for $t = t_i + (N-1)T_s$, the terms in (12) can be rewritten as $V_{i|j}(x_t) = x_i^T[N-1]P_{i|j}x_i[N-1] = (\bar{x}_i[N-1])^T \Theta(P_{i|j})$ and

$$\int_t^{t+T_s} x_\tau^T Q_{K_i} x_\tau \, d\tau = W_i[N] - W_i[N-1],$$

respectively, where $\Theta(X)$ is the invertible map from an $n \times n$ symmetric matrix X to a column vector in $\mathbb{R}^{n(n+1)/2}$; the column vector is made by stacking the upper triangular parts of X on top of one another, where the off-diagonal terms are doubled. Substituting all the equations into (12) yields

$$\bar{x}_i^T[N-1]\Theta(P_{i|j+1}) = \Delta W_i[N-1] + \bar{x}_i^T[N]\Theta(P_{i|j}). \tag{43}$$

where $\Delta W_i[N-1] := W_i[N] - W_i[N-1]$. Note that (43) holds for all $N = 1, 2, \cdots, N_{i,max}$. Collecting them for $N = 1, 2, \cdots, N_{i,max}$, one obtains the matrix form of (43) as

$$\bar{X}_{i,0}^T \Theta(P_{i|j+1}) = \Delta \bar{W}_i + \bar{X}_{i,1}^T \Theta(P_{i|j}), \tag{44}$$

where $\Delta \bar{W}_i := \left[\Delta W_i[0] \;\; \Delta W_i[1] \;\; \cdots \;\; \Delta W_i[N_{i,max}]\right]^T \in \mathbb{R}^{N_{i,max}}$. Therefore, under Assumption 3, the exact $P_{i|j+1}$ satisfying (12) can be obtained by solving the batch LS equation:

$$\Theta(P_{i|j+1}) = F_i\left[\Delta \bar{W}_i\right] + G_i\left[\Theta(P_{i|j})\right], \tag{45}$$

where F_i and G_i are the matrices defined, respectively, as $F_i := (\bar{X}_{i,0}\bar{X}_{i,0}^T)^{-1}\bar{X}_{i,0}$ and $G_i := F_i\bar{X}_{i,1}^T$. Now, by recursively solving (45) k-times with $P_{i|0} = P_i$, one can obtain $P_{i+1} = P_{i|k}$ at each policy evaluation step. In the procedures, *what all the agent should additionally do at $(j+1)$-th sub-step is the substitution of $\Theta(P_{i|j})$ evaluated at j-th sub-step into* (45).

6.2 Determination of the iteration horizon k

To guarantee the stability and convergence, the iteration horizon k and/or the update horizon $\hbar$ should be cautiously decided at each iteration. For example, $\hbar$ should be sufficiently

small for VI-mode of convergence by Theorem 7. In this subsection, we mainly concentrate on the determination of k at each iteration under initial stabilizing policy u_0.

Notice that all the stability and monotone convergence conditions shown in the Theorems in Section 5 can be easily checked as long as the agent knows $\mathscr{L}(K_i, P_{i+1})$ *a priori*, which depends highly on k and converges to 0 as $k \to \infty$ if K_i is stabilizing. However, evaluating $\mathscr{L}(K_i, P_{i+1})$ is not a trivial task if the system matrix A is not known (see (24) and the definition of $\mathscr{L}(K,P)$ and $\mathscr{R}(P)$). Instead, if one has $\bar{X}_{i,0}$, $\bar{X}_{i,1}$, and the information of $\mathscr{L}(K_i, P_{i|j})$ *a priori*, $\mathscr{L}(K_i, P_{i|j+1})$ can be obtained from the equation

$$x_t^T \mathscr{L}(K_i, P_{i|j+1}) x_t = x_{t+T_s}^T \mathscr{L}(K_i, P_{i|j}) x_{t+T_s} \tag{46}$$

which can be derived using (16) in Lemma 3 and substituting $e^{A_i T_s} x_t = x_{t+T_s}$. Since (46) holds for all $t \in [t_i, t_i + NT_s)$, similarly to (43), one can represent (46) as

$$\bar{x}_i^T [N-1] \Theta(\mathscr{L}(K_i, P_{i|j+1})) = \bar{x}_i^T [N] \Theta(\mathscr{L}(K_i, P_{i|j})).$$

Then, by the same procedure to (44)–(45), we can derived the following LS equation

$$\Theta(\mathscr{L}(K_i, P_{i|j+1})) = G_i \big[\Theta(\mathscr{L}(K_i, P_{i|j}))\big], \tag{47}$$

which uniquely determines $\mathscr{L}(K_i, P_{i|j+1})$ under Assumption 3. Moreover, by the recursive application to $\mathscr{L}(K_i, P_{i|j+p})$ ($p \in \mathbb{N}$), the result can be easily generalized as follows:

$$\Theta(\mathscr{L}(K_i, P_{i|j+p})) = (G_i)^p \big[\Theta(\mathscr{L}(K_i, P_{i|j}))\big]. \tag{48}$$

Therefore, if $\mathscr{L}(K_i, P_i)$ ($P_i = P_{i|0}$) is given *a priori*, the agent can evaluate $\mathscr{L}(K_i, P_{i+1})$ ($P_{i+1} = P_{i|k}$) for any k by recursively solving (47) and/or (48). This allows the agent to determine k by checking the stability and monotone convergence conditions online at each iteration without knowing the matrix A. For example, one can consider the inequality

$$\mathscr{L}(K_i, P_{i|j}) \le Q_{K_{i|j}} \tag{49}$$

for the closed-loop stability, where $K_{i|j} := R^{-1} B^T P_{i|j}$, and determine k by $k = j$ if $\mathscr{L}(K_i, P_{i|j})$ satisfies (49) (see Theorem 3). If A_i is Hurwitz, there is $k \in \mathbb{N}$ such that the inequality (49) holds by the convergence of $\mathscr{L}(K_i, P_{i|j})$ to zero (see Theorem 2). Therefore, one can guarantee the closed-loop stability under an initial stabilizing policy by determining k at every iteration in such a way that (49) holds for the respective i-th iteration.

On the other hand, the prior information of $\mathscr{L}(K_i, P_i)$ can be easily obtained when $P_0 = 0$. In this case, we have $\mathscr{L}(K_0, P_0) = Q_{K_0}$ which is *not* connected to the matrix A. Then, $\mathscr{L}(K_i, P_i)$ can be obtained by solving (47) or (48) and then using (24), which yields the sequence $\mathscr{R}(P_1)$, $\mathscr{R}(P_2)$, $\mathscr{R}(P_3)$, $\cdots$. For an arbitrary P_0, however, this cannot be done due to the lack of knowledge of the matrix A. In this case, instead of imposing the assumption that A is known *a priori*, we can use another stability condition derived from (28) in Theorem 4 as shown below.

If defining $Q_{i|0} := Q_{K_i}$ and $Q_{i|j} := e^{A_i^T T_s} Q_{i|j-1} e^{A_i T_s}$, then, similarly to (46)–(48), one obtains the following LS equation

$$\Theta(Q_{i|j+p}) = (G_i)^p \big[\Theta(Q_{i|j})\big] \tag{50}$$

for any $j, p \in \mathbb{Z}_+$ and hence, any information about the system matrix A and $\mathscr{L}(K_i, P_i)$ is not necessary for the computation of $Q_{i|j+p}$. Substituting $Q_{i|j}$ and $Q_{K_{i|j}}$ into (28), we obtain the stability condition "$Q_{i|j} \le Q_{K_{i|j}}$", which can be checked by solving (50), without knowing the matrix A. By Theorem 4 and the above discussions, the choice $k = j$ for j satisfying $Q_{i|j} \le Q_{K_{i|j}}$ at every i-th step preserves the stability of the closed-loop system.

7 Simulations

To verify the performance of the proposed implementation methods and further discuss the properties of I-GPI shown in this paper, we simulated I-GPI (Algorithm 1) with the following LQR problem for the load frequency control system:

$$\begin{cases} A = \begin{bmatrix} -5 & 0 & -4 \\ 2 & -2 & 0 \\ 0 & 0.1 & -0.08 \end{bmatrix}, \ B = \begin{bmatrix} 0 \\ 0 \\ -0.1 \end{bmatrix}, \\ Q = \mathrm{diag}\{20, 10, 5\}, \ R = 0.15 \end{cases} \tag{51}$$

which is the same framework given by Saadat (2002, Example 12.11), except that the governor speed regulation was set to 1.25 per unit. In all the simulations, the sine-wave exploration $u_t = 10^{-2} \sin 20\pi t$ was applied during T_s [s] before every policy evaluation step, and 12 data pairs $(x_i[N], W_i[N])$ were collected at each iteration for the data-driven implementations, *i.e.*, $N_{i,max} = 12 \ \forall i \in \mathbb{Z}_+$. Therefore, P_i and K_i were updated every $(1+12)T_s$ [s] in the simulations. In all the simulations, P_i was updated by the LS equation (45), and for each simulation the iteration horizon k was either set to a fixed value (Simulations 1–4) or determined by the LS method presented in Section 6.2 (Simulation 5).

Simulation 1: This simulation is intended to verify that all I-GPI algorithms with the same update horizon $\hbar$ yield the same sequence $\{P_i\}$ (see Theorem 1 and Remark 1). The simulation was performed with $(P_0, K_0) = (0, 0)$ for the same $\hbar = 0.3$ [s] and several different iteration horizons $k = 3, 6, 12, \cdots$. The simulation results are shown in Fig. 3, where the time axes were superposed and drawn only for the case of $k = 3, 12$. Note that all the sampling period T_s was set by the equation $kT_s = \hbar = 0.3$ [s], so the simulation results have different scales in the time domain. On the other hand, one can see from Fig. 3 that all the I-GPI algorithms with the same $\hbar$ yield the same sequence $\{P_i\}$, verifying in the

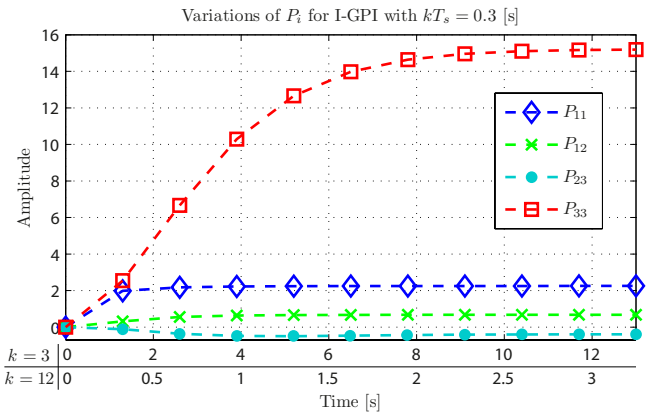


Fig. 3. (**Simulation 1**) Variations of P_i for I-GPI with $\hbar = 0.3$ [s].

iteration domain the equivalence of all the I-GPI methods that have the same $\hbar$.

Simulation 2 (PI-mode of convergence): In this simulation, the initial conditions P_0 and K_0 were set to $P_0 = \text{diag}\{10, 10, 20\}$ and $K_0 = \begin{bmatrix} 0 & 0 & -14 \end{bmatrix}$, respectively, so that the initial pair (P_0, K_0) satisfies $\mathscr{L}(K_0, P_0) \leq 0$. Fig. 4 shows the simulation results for $k = 5$ and $T_s = 20$ [ms]. In this case, as stated in Theorem 5 and can be seen from Fig. 4, all the closed-loop systems are stable and P_i monotonically converges to P_{K^*} in PI-mode. Fig. 4(a) illustrates the state trajectories where the marked points indicate the time instant the policy was changed by the I-GPI agent. Here, the states rapidly vary right after the marked points due to the exploration $u_t = 10^{-2} \sin 20\pi t$ applied before every policy evaluation of I-GPI. From this figure, one can see that the states remain in a small bounded region by the stability argument. In addition, Fig. 4(b) shows the convergence of P_i to P_{K^*}, where the diagonals (P_{11} and P_{33}) are monotonically decreasing. This PI-mode of convergence becomes obvious by Fig. 4(c), which shows the eigenvalues of the difference $P_i - P_{i-1}$ are always negative, implying monotone decreasing $0 \leq P_i \leq P_{i-1} \leq \cdots \leq P_0$. Therefore, Fig. 4(b) and (c) show PI-mode of convergence stated in Theorem 5.

Simulation 3 (VI-mode of convergence): To further investigate the monotone convergence in VI-mode, an additional simulation was performed with the same settings as in Simulation 1, except that $\hbar$ was given by $\hbar = 1.2$ [s]. Then, the results were compared with those of Simulation 1 ($\hbar = 0.3$ [s]) as shown in Fig. 5 and Table 1. In both simulations, T_s was set to $T_s = 0.1$ [s]. Fig. 5 shows the variations of eigenvalues of $\mathscr{L}(K_i, P_i)$. In Fig. 5(a), it is shown that all the eigenvalues of $\mathscr{L}(K_i, P_i)$ remain positive for $\hbar = 0.3$ [s], implying VI-mode of convergence by Theorem 7. Here, the convergence to P_{K^*} is verified by Fig. 3 and the monotonicity can be seen from Table 1, where the minimum eigenvalues of $P_i - P_{i-1}$ for $\hbar = 0.3$ [s] are all positive. This implies (40) with $l \to \infty$ in Theorem 7. On the other hand, in the case of $\hbar = 1.2$ [s], only the minimum eigenvalue of $P_1 - P_0$ ($i = 1$) is positive due to the initial condition $\mathscr{L}(K_0, P_0) \geq 0$, but the others are not due to violations of $\mathscr{L}(K_i, P_i) \geq 0$ for $i \geq 1$,

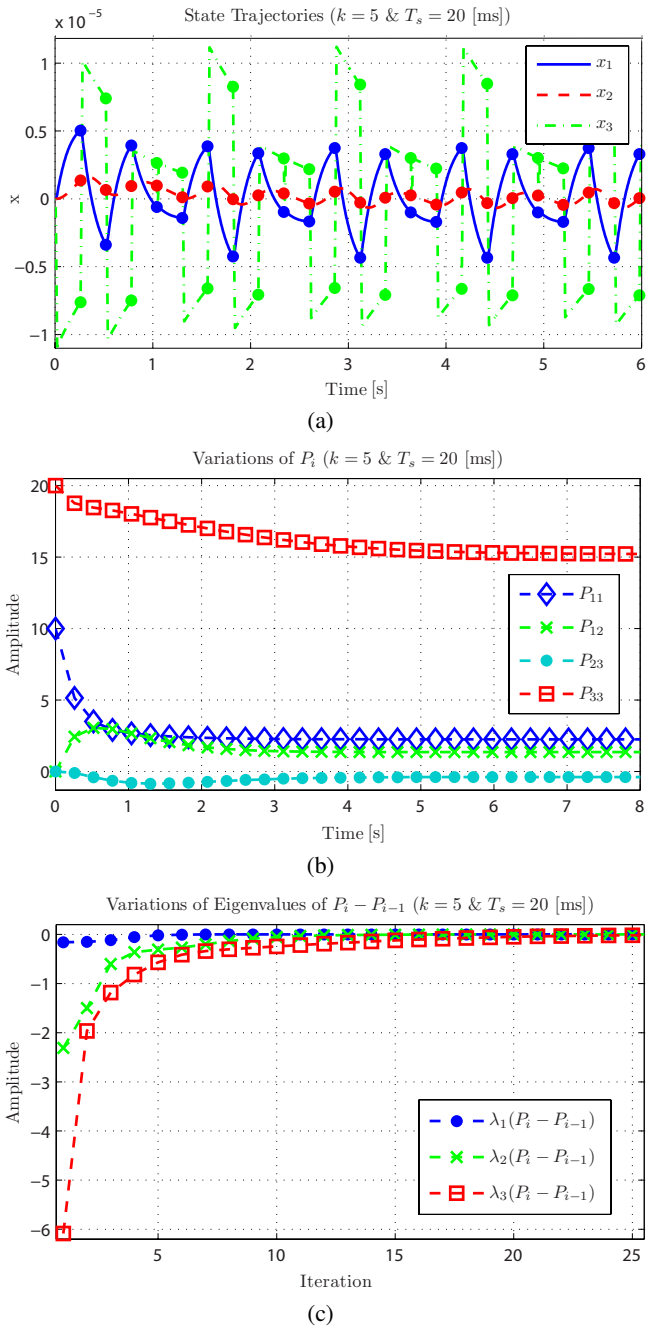


Fig. 4. (**Simulation 2**) Variations of (a) state variable x, (b) P_i, and (c) eigenvalues of the difference $P_i - P_{i-1}$ for the I-GPI with $k = 5$ and $T_s = 20$ [ms]; the initial conditions are given by $P_0 = \text{diag}\{10, 10, 20\}$ and $u_0 = 14x_3$.

as shown in Fig. 5(b) and Table 1 for $\hbar = 1.2$ [s]. Therefore, while P_i for $\hbar = 1.2$ [s] is actually shown to converge to P_{K^*} ($\because \mathscr{L}(K_i, P_i) \to 0$ by Fig. 5(b)), unlike the case with the small $\hbar = 0.3$ [s], the convergence is not monotone for this relatively large update horizon $\hbar = 1.2$ [s].

Simulation 4 (Local monotone convergence): The next simulation focuses on the local monotone convergence re-

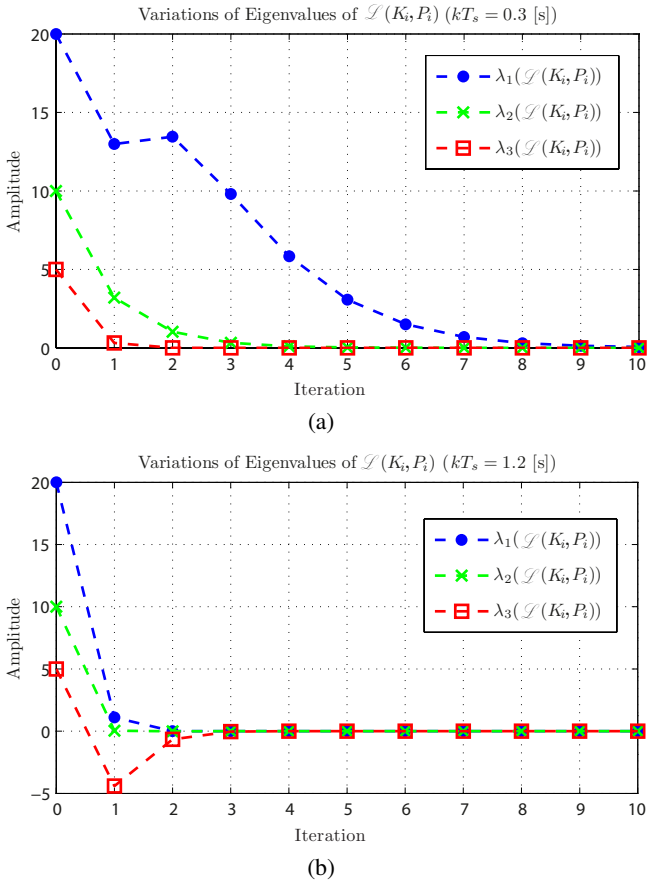


Fig. 5. (**Simulation 3**) Variations of $\mathscr{L}(K_i,P_i)$ for (a) $\hbar=0.3$ [s] and (b) $\hbar=1.2$ [s].

gions Ω_i^1 and Ω_i^2 in Theorem 6. The I-GPI simulations were performed for various $\hbar>0$, as shown in Fig. 6 and Table 2. The time horizon was fixed to $T_s=0.1$ [s] for all the simulations, and all of the other parameters were set to those in Simulation 1. For each iteration and each $\hbar$, we verified that A_i remains Hurwitz. Table 2 shows the variations of $\rho_{(1,i)}+\|e^{A_i\hbar}\|$ for various $\hbar$. Note that from the definition of Ω_i^1, one can see that if $\mathscr{L}(K_i,P_i)$ satisfies "$\rho_{(1,i)}+\|e^{A_i\hbar}\|<1$", then P_i belongs to Ω_i^1, so that local 1st-order monotone decreasing is guaranteed by Theorem 6. As shown in Table 2, the quantities are less than "1" for $i\geq 2$ and $h\geq 1.2$ [s]. For $h\geq 0.3$ [s], $P_i\in\Omega_i^1$ for $i\geq 5$ can be inferred from Table 2. Table 2 also shows that i) $\rho_{(1,i)}+\|e^{A_i\hbar}\|$ converges to the fixed values as $i\to\infty$ for all $\hbar$, and ii) the

Table 1
(Simulation 3) Variations of the min. eigenvalue $\lambda_3(P_i-P_{i-1})$

i	$\hbar=0.3$[s]	$\hbar=1.2$[s]	i	$\hbar=0.3$[s]	$\hbar=1.2$[s]
1	1.16e-00	1.59e-00	6	6.94e-08	-7.59e-06
2	3.53e-02	-2.30e-00	7	3.36e-09	-8.98e-08
3	1.07e-03	-2.76e-01	8	1.66e-10	-4.80e-10
4	3.72e-05	-1.38e-02	9	8.29e-12	-2.53e-11
5	1.52e-06	-3.97e-04	10	4.74e-13	-2.22e-12

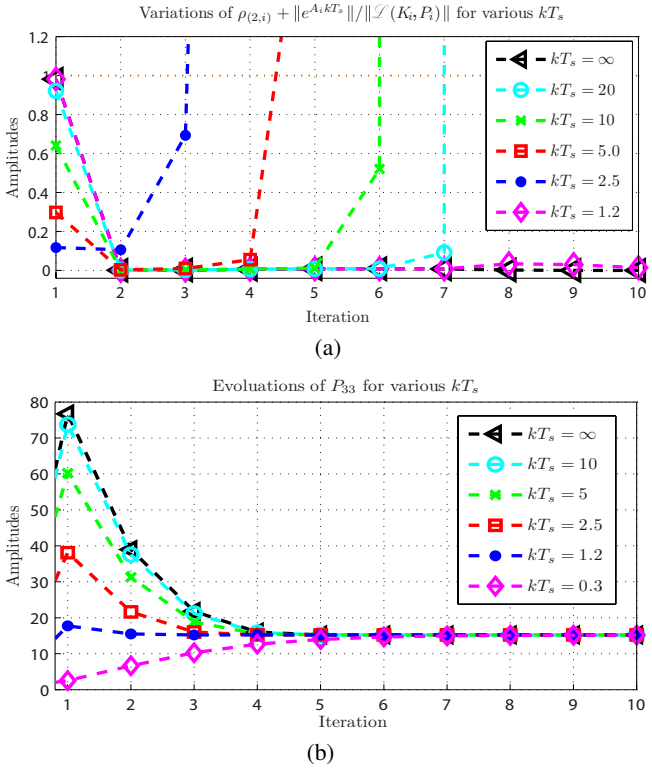


Fig. 6. (**Simulation 4**) Variations of (a) $\rho_{(2,i)}+\|e^{A_i\hbar}\|/\|\mathscr{L}(K_i,P_i)\|$ and (b) P_{33} for several $\hbar$.

fixed values in the limit also converge to zero as $\hbar\to\infty$, so condition $P_i\in\Omega_i^1$ is more likely satisfied as $\hbar\to\infty$.

Fig. 6(a) shows the variations of $\rho_{(2,i)}+\|e^{A_i\hbar}\|/\|\mathscr{L}(K_i,P_i)\|$. Remember that if this quantity is less than '1', then local 2nd-order monotone decreasing condition $P_i\in\Omega_i^2$ in Theorem 6 is achieved. Unlike the results in Table 2, this quantity becomes larger than '1' after finite number of iterations for $\hbar\leq 10$ [s] as shown in Fig. 6(a), *e.g.*, $P_i\notin\Omega_i^2$ for $i\geq 5$ and $\hbar=2.5$ [s]. On the other hand, for $\hbar=20$ [s] and $\hbar=\infty$ [s] (I-PI), one can infer from Fig. 6(a) that $P_i\in\Omega_i^2$ for all i, so 2nd-order monotone convergence is achieved by Theorem 6. This corresponds to the fact that I-GPI with sufficiently large $\hbar$ ($\hbar\geq 20$ [s] in this case) approximately equals I-PI which guarantees monotone convergence to P_{K^*} with order 2. As implied by Theorem 6 and the above results, the parameters P_i converge to the optimal solution as shown in Fig. 6(b) for the parameter P_{33} and for all $\hbar$.

Simulation 5 (I-GPI with adaptive iteration horizon k): The purpose of this last simulation is to verify the performance of I-GPI when the iteration horizon k is determined by the method presented in Section 6.2. All the simulation conditions were set to those used in Simulation 1, but k is determined based on the inequality $Q_{i|j}\leq Q_{K_{i|j}}$, where $Q_{i|j}$ is evaluated by (50). At each i-th iteration, the agent recursively evaluates $Q_{i|j}$ and $P_{i|j}$ and chooses $k=j$ if $Q_{i|j}\leq Q_{K_{i|j}}$. In this simulation, the inequality is checked at every even

Table 2
(Simulation 4) The variations of $\rho_{(1,i)} + \|e^{A_i\hbar}\|$ for several $\hbar$.

i	update horizon $\hbar$					
	0.3 [s]	1.2 [s]	2.5 [s]	5.0 [s]	10 [s]	∞ [s]
1	1.2030	2.3497	5.9446	12.8189	18.4420	19.6395
2	1.2084	0.4684	0.2321	0.2366	0.2414	0.2421
3	1.1056	0.4569	0.1725	0.1844	0.2045	0.2078
4	1.0174	0.4562	0.1161	0.0753	0.1064	0.1129
5	0.9627	0.4562	0.1049	0.0092	0.0162	0.0188
6	0.9335	0.4562	0.1046	0.0027	0.0003	0.0003
∞	0.9072	0.4562	0.1046	0.0026	0.0000	0.0000

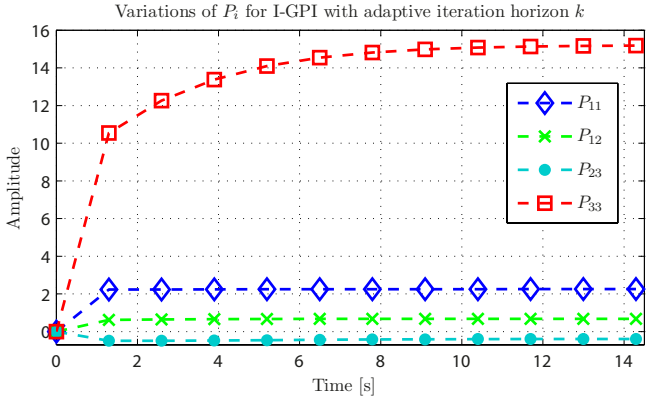


Fig. 7. (**Simulation 5**) Variations of P_i for I-GPI with adaptive k.

Table 3
(Simulation 5) Variations of k and the maximum of $\mathrm{Re}\lambda_j(A_i)$

i	k	$\max_j \mathrm{Re}\lambda_j(A_i)$	i	k	$\max_j \mathrm{Re}\lambda_j(A_i)$
1	8	-0.1705	5	2	-1.4981
2	2	-0.1705	6	2	-1.5216
3	2	-1.4065	7	2	-1.5364
4	2	-1.4616	8	2	-1.5454

number of $j = 2,4,6,\cdots$ until it is satisfied. The simulation results are shown in Fig. 7 and Table 3, where Fig. 7 describes the convergence of P_i to P_{K^*}. In this case, the iteration horizon k was determined to $k = 8$ for $i = 1$ and $k = 2$ otherwise, all by $Q_{i|j} \leq Q_{K_{i|j}}$. This preserves the stability of A_i as shown in Table 3.

8 Conclusions

This paper focused on I-GPI methods applied to LQR problems and investigated their properties. We have shown that i) I-GPIs with the same update horizon $\hbar = kT_s$ are all equal and have the same convergence speed in the iteration domain, and ii) the approximated value function in policy evaluation of I-GPI monotonically converges to the exact one as $\hbar \to \infty$. These implied that i) I-GPI in the limit $\hbar \to \infty$ is the same as I-PI, and ii) for the same $\hbar$, a trade-off exists between the computational complexity k and the update speed T_s. Based on these results, a pair of monotone convergence properties, namely, PI- and VI-mode of convergence, were investigated for I-GPI (Fig. 2(a)). In PI-mode, I-GPI behaves like I-PI ($\hbar \to \infty$), and in VI-mode, it performs like VI for DT LQR and infinitesimal GPI ($\hbar \to 0$). Taking all of these into consideration, a new spectral classification of I-RL methods was established with respect to $\hbar$, where infinitesimal GPI and I-PI are posed on two extreme tips of the spectrum (Fig. 1). Two matrix inequality conditions for stability and the region of local monotone convergence were also presented with detailed discussions in relation to PI-mode of convergence and stability (Fig. 2(b)). LS implementation methods of I-GPI were also proposed, which are able to adaptively determine the iteration horizon k based on those properties. Finally, five numerical simulations were carried out to verify and further investigate those individual properties and implementation methods.

Acknowledgements

The authors appreciate the Associate Editor and anonymous reviewers for their valuable suggestions.

References

Al-Tamimi, A. (2007) *Discrete-time control algorithms and adaptive intelligent systems designs. Ph. D thesis*, TX, USA: University of Texas at Arlington.

Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K. G., Lewis, F. L., & Dixon, W. E. (2013) A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems, *Automatica, 49*(1), 82–92.

Bertsekas, D. P. & Tsitsiklis, J. N. (1996) *Neuro-dynamic programming.* Belmont, MA: Athena Scientific.

Doya, K. (2000) Reinforcement learning in continuous-time and space. *Neural Computation, 12*, 219–245.

Feitzinger, F., Hylla, T., & Sachs, E. W. (2009) Inexact Kleinman-Newton method for Riccati equations. *SIAM Journal on Matrix Analysis & App., 31*(2), 272–288.

Jiang, Y. & Jiang, Z. (2010) Approximate dynamic programming for output feedback control In *Proc. Chinese Control Conference*, 5815–5820.

Hanselmann, T., Noakes, L., & Zaknich, A. (2007) Continuous-time adaptive critics. *IEEE Trans. Neural Networks, 18*(3), 631–647.

Hewer, G. (1971) An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Trans. Automatic Control, 16*(4), 382–384.

Khalil, H. K. (2002) *Nonlinear Systems.* Prentice Hall.

Kleinman, D. (1968) On the iterative technique for Riccati equation computations. *IEEE Trans. Automatic Control, 13*(1), 114–115.

Lancaster, P. & Rodman, L. (1995). *Algebraic Riccati equations*. Oxford University Pres, New York.

Lendelius, T. (1997) *Reinforcement learning and distributed local model synthesis. Ph. D. thesis* Sweden: Linkoping University.

Lee, J. Y., Park, J. B., & Choi, Y. H. (2010). A novel generalized value iteration scheme for uncertain continuous-time linear systems. In *Proc. CDC*, 4637–4642.

Lee, J. Y., Park, J. B., & Choi, Y. H. (2011). On generalized policy iteration for continuous-time linear systems. In *Proc. CDC*, 1722–1728.

Lee, J. Y., Park, J. B., & Choi, Y. H. (2012) Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems. *Automatica, 48*(11), 2850–2859.

Lewis, F. L. & Vrabie, D. (2009) Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE circuits and systems magazine, 9*(3), 32–50.

Prokhorov, D. V. & Wunsch II, D. C. (1997) Adaptive critic designs. *IEEE Trans. Neural Networks*, 8(5), 997–1007.

Puterman, M. L. & Shin, M. C. (1978) Modified policy iteration algorithms for discounted Markov decision problems. *Management science*, 24, 1127–1137.

Saadat, H. (2002) *Power system analysis*. McGraw-Hill Primis Custom Publishing.

Si, J., Barto, A. G., Powell, W. B., & Wunsch, D. (2004) *Handbook of learning and approximate dynamic programming*. Wiley-IEEE Press.

Stoorvogel, A. A. & Weeren, Arie J. T. M. The discrete-time Riccati equation related to the H_∞ control problems. *IEEE Trans. Automatic Control, 39*(5), 686–691.

Sutton, R. S. & Barto, A. G. (1998) *Reinforcement learning–an introduction*. MIT Press, Cambridge, Massachussetts.

van Nunen, J. A. E. E. (1976) A set of successive approximation methods for discounted Markovian decision problems. *Mathematical Methods of Operations Research, 20*(5), 203—208.

Vamvoudakis, K. G. & Lewis, F. L., (2010) Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem, *Automatica, 46*(5), 878–888.

Vrabie, D. (2009) *Online adaptive optimal control for continuous-time systems. Ph. D. thesis*, TX, USA: University of Texas at Arlington.

Vrabie, D. & Lewis, F., L. Generalized policy iteration for continuous-time systems. In *Proc. Int. Joint Conf. Neural Networks*, Atlnata, GA, USA, 3224–3231, 2009.

Wang D., Liu D., Wei Q., Zhao D., & Jin N. Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming. *Automatica, 48*(8), 1825–1832.

Wang, F., Y., Zhang, H., & Liu, D. Adaptive dynamic programming: an introduction. *IEEE Computational Intelligence Magazine, 4*(3), 39–47.

Webos, P., J. (1992) Approximate dynamic programming for real-time control and neural modeling. *Handbook of Intelligent Control,* D. A. White and D.A. Sofge, Eds. New York: Van Nostrand Reinhold.

Zhang, H., Huang, J., & Lewis, F. L. (2009) An improved method in receding horizon control with updating of terminal cost function. *Applications of Intelligent Control to Engineering Systems, Springer,* 365–393.

Appendix

A Proof of Lemma 2

Assume that $\{P_i\}$ converges. Then, $\{K_i\}$ also converges due to $K_i = R^{-1}B^T P_i$ for $i \geq 1$. Let $\bar{K}$ and $\bar{P}$ be their respective limit points. Then, taking the limit $i \to \infty$ of (22), we have

$$0 = \lim_{i\to\infty}(P_{i+1} - P_i) = \lim_{i\to\infty} \int_0^\hbar e^{A_i^T \tau} \mathscr{L}(K_i, P_i) e^{A_i \tau}\, d\tau,$$

which implies $\mathscr{L}(\bar{K}, \bar{P}) = 0$. Since $\mathscr{L}(K_i, P_i) = \mathscr{R}(P_i)$ holds for $i \geq 1$ by $K_i = R^{-1}B^T P_i$, "$\mathscr{L}(\bar{K}, \bar{P}) = 0$" again implies the ARE "$\mathscr{R}(\bar{P}) = 0$". This ARE has the unique solution $\bar{P} = P_{K^*}$ since $(A, B, S^{1/2})$ is stabilizable and detectable. Therefore, the limit points satisfy $\bar{P} = P_{K^*}$ and $\bar{K} = K^*$, completing the proof.

B Proof of Lemma 5

The proof is almost parallel to that given by Feitzinger *et al.* (2009, Theorem 4.3). First, note that $\mathscr{R}(P_{i+1}) = \mathscr{L}(K_{i+1}, P_{i+1})$ and (5) in Lemma 1 allow the following expression:

$$\underbrace{A_{i+1}^T P_{i+1} + P_{i+1} A_{i+1} + Q_{K_{i+1}}}_{=\mathscr{L}(K_{i+1}, P_{i+1})} - \mathscr{L}(K_i, P_{i+1}) = -\Delta K_i^T R \Delta K_i.$$

Assuming $A_{i+1}x = \lambda x$ for $\lambda \in \mathbb{C}$ with $\mathrm{Re}(\lambda) \geq 0$ and $x \in \mathbb{C}^n$ with $x \neq 0$, we have

$$\bar{x}^T \left[(\bar{\lambda} + \lambda) P_{i+1} + Q_{K_{i+1}} - \mathscr{L}(K_i, P_{i+1}) \right] x = -\bar{x}^T \Delta K_i^T R \Delta K_i x,$$

where $\bar{x}$ and $\bar{\lambda}$ are the complex conjugates of x and λ, respectively. Here, by the assumptions of $0 \leq P_{i+1}$ and $\mathscr{L}(K_i, P_{i+1}) \leq Q_{K_{i+1}}$, the matrix on the left hand side is positive semi-definite, but $-\Delta K_i^T R \Delta K_i$ on the right hand side is obviously negative semi-definite. Therefore, we obtain $\bar{x}^T (\Delta K_i^T R \Delta K_i) x = 0$. This again implies $(K_i - K_{i+1})x = 0$ due to the positive definiteness of R, and thereby we finally obtain $K_i x = K_{i+1} x$, that is, $A_i x = A_{i+1} x$. This implies that K_i is not a stabilizing policy (since $\mathrm{Re}(\lambda) \geq 0$), but K_i is assumed stabilizing in Lemma 5, a contradiction. Therefore, K_{i+1} should also be a stabilizing policy, which completes the proof.