Control Theory and Technology (2023) 21:374–389
https://doi.org/10.1007/s11768-023-00163-w

RESEARCH ARTICLE

A Lyapunov characterization of robust policy optimization

Leilei Cui[1] · Zhong-Ping Jiang[1]

Received: 16 May 2023 / Revised: 26 June 2023 / Accepted: 27 June 2023 / Published online: 22 September 2023

Abstract

In this paper, we study the robustness property of policy optimization (particularly Gauss–Newton gradient descent algorithm which is equivalent to the policy iteration in reinforcement learning) subject to noise at each iteration. By invoking the concept of input-to-state stability and utilizing Lyapunov's direct method, it is shown that, if the noise is sufficiently small, the policy iteration algorithm converges to a small neighborhood of the optimal solution even in the presence of noise at each iteration. Explicit expressions of the upperbound on the noise and the size of the neighborhood to which the policies ultimately converge are provided. Based on Willems' fundamental lemma, a learning-based policy iteration algorithm is proposed. The persistent excitation condition can be readily guaranteed by checking the rank of the Hankel matrix related to an exploration signal. The robustness of the learning-based policy iteration to measurement noise and unknown system disturbances is theoretically demonstrated by the input-to-state stability of the policy iteration. Several numerical simulations are conducted to demonstrate the efficacy of the proposed method.

Keywords Policy optimization · Policy iteration (PI) · Input-to-state stability (ISS) · Lyapunov's direct method

1 Introduction

Through reinforcement learning (RL) techniques, agents can iteratively minimize the specific cost function by interacting continuously with unknown environment. Policy optimization is fundamental for the development of RL algorithms as introduced in [1]. Policy optimization first parameterizes the control policy, and then, the performance of the control policy is iteratively improved by updating the parameters along the gradient descent direction of the given cost function. Since the linear quadratic regulator (LQR) problem is tractable and widely applied in many engineering fields, it provides an ideal benchmark example for the theoretical analysis of policy optimization. For the LQR problem, the control policy is parameterized by a control gain matrix, and the gradient of the quadratic cost with respect to the control gain is associated with a Lyapunov matrix equation. Based on these results, various policy optimization algorithms, including vanilla gradient descent, natural gradient descent and Gauss–Newton methods, are developed in [2–5]. Compared with other policy optimization algorithms with a linear convergence rate, the control policies generated by the Gauss–Newton method converge quadratically to the optimal solution.

It is noticed that the Gauss–Newton method with the step size of $1/2$ coincides with the policy iteration (PI) algorithm [6, 7], which is an important iterative algorithm in RL and adaptive/approximate dynamic programming (ADP) [1, 8, 9]. From the perspective of the PI, the Lyapunov matrix equation for computing the gradient can be considered as policy evaluation. The update of the policy along the gradient direction can be interpreted as policy improvement. The steps of policy evaluation and policy improvement are iterated in turn to find the optimal solution of LQR. Various PI algorithms have been proposed for important classes of linear/nonlinear/time-delay/time-varying systems for optimal stabilization and output tracking [10–14]. In addition, PI has been successfully applied to sensory motor control [15], and autonomous driving [16, 17].

This work was supported in part by the National Science Foundation (Nos. ECCS-2210320, CNS-2148304).

✉ Leilei Cui
l.cui@nyu.edu

Zhong-Ping Jiang
zjiang@nyu.edu

[1] Department of Electrical and Computer Engineering, New York University, Brooklyn, NY 11201, USA

The convergence of the PI algorithm is ensured under the assumption that the accurate knowledge of system model is accessible. However, in reality, the system model obtained by system identification [18] is used for the PI algorithm or the PI algorithm is directly implemented through a data-driven approach using input-state data [10, 19–22]. Consequently,the PI algorithm is hardly implemented accurately due to modeling errors, inaccurate state estimation, measurement noise, and unknown system disturbances. The robustness of the PI algorithm to unavoidable noise is an important property to be investigated, which lays a foundation for better understanding RL algorithms. There are several challenges for studying the robustness of the PI algorithm. Firstly, the nonlinearity of the PI algorithm makes it hard to analyze the convergence property. Secondly, it is difficult to quantify the influence of noise, since noise may destroy the monotonic property of the PI algorithm or even result in a destabilizing controller.

In this paper, we study the robustness of the PI algorithm in the presence of noise. The contributions are summarized as follows. Firstly, by viewing the PI algorithm as a nonlinear system and invoking the concept of input-to-state stability (ISS) [23], particularly the small-disturbance ISS [24, 25], we investigate the robustness of the PI algorithm under the influence of noise. It is demonstrated that when subject to noise, the control policies generated by the PI algorithm will eventually converge to a small neighborhood of the optimal solution of LQR as long as noise is sufficiently small. Different from [24, 25], where the analysis is trajectory-based, we directly utilize Lypuanov's direct method to analyze the convergence of the PI algorithm under disturbances. As a result, an explicit expression of the upperbound on the noise is provided. The size of the neighborhood in which the control policies will ultimately stay is demonstrated as a quadratic function of the noise. Secondly, by utilizing Willems' fundamental lemma, a learning-based PI algorithm is proposed. Compared with the conventional learning-based control approach where the exploratory control input is hard to design such that the persistent excitation condition is satisfied [24], the persistently exciting exploratory signal of the proposed method can be easily designed by checking the rank condition of a Hankel matrix related to the exploration signal. Finally, based on the small-disturbance ISS property of the PI algorithm, we demonstrated that the proposed learning-based PI algorithm is robust to the state measurement noise and unknown system disturbances.

The remaining contents of the paper are organized as follows. Section 2 reviews the LQR problem and the celebrated PI algorithm. In Sect. 3, the small-disturbance ISS property of the PI algorithm is studied. Section 4 proposes a learning-based PI algorithm and the robustness of the algorithm is analyzed. Several numerical examples are given in Sect. 5, followed by some concluding remarks in Sect. 6.

Notations In this paper, $\mathbb{R}$ ($\mathbb{R}_+$) denotes the set of (nonnegative) real numbers, $\mathbb{Z}_+$ denotes the set of nonnegative integers, and $\mathbb{S}^n$ denotes the set of n-dimensional real symmetric matrices. $|\cdot|$ denotes the Euclidean norm for a vector. $\|\cdot\|$ denotes the spectral norm and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. $\|\cdot\|_\infty$ denotes the ℓ^∞-norm, that is $\|A\|_\infty = \sup_{i\in\mathbb{Z}_+}\|A_i\|$ for $A = \{A_i\}_{i=0}^\infty$. $\bar{\lambda}(\cdot)$ and $\underline{\lambda}(\cdot)$ denote the maximum and minimum eigenvalues of a real symmetric matrix, respectively. $\mathrm{Tr}\,(\cdot)$ denotes the trace of a matrix. A continuous function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ is a $\mathcal{K}$-function if it is strictly increasing and vanishes at zero. A function $\beta(r,t) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is a $\mathcal{KL}$-function if for each fixed t, $\beta(\cdot,t)$ is a $\mathcal{K}$-function and for each fixed r, $\beta(r,t)$ tends to zero as $t \to \infty$. For $Z \in \mathbb{R}^{m\times n}$ and $U \in \mathbb{S}^{m+n}$, define $\mathcal{H}(U,Z)$ as

$$\mathcal{H}(U,Z) = \begin{bmatrix} I_n & -Z^{\mathrm{T}} \end{bmatrix} U \begin{bmatrix} I_n \\ -Z \end{bmatrix}.$$

2 Preliminaries and problem formulation

2.1 Policy iteration for discrete-time LQR

The discrete-time linear time-invariant (LTI) system is represented as

$$x_{k+1} = Ax_k + Bu_k. \tag{1}$$

where $x_k \in \mathbb{R}^n$ and $u_k \in \mathbb{R}^m$ are the state and control input, respectively; A and B are system matrices with compatible dimensions.

Assumption 1 The pair (A, B) is controllable.

Under Assumption 1, the discrete-time LQR is to minimize the following accumulative quadratic cost

$$J_d(x_0, u) = \sum_{k=0}^{\infty} x_k^{\mathrm{T}} Q x_k + u_k^{\mathrm{T}} R u_k, \tag{2}$$

where $Q = Q^{\mathrm{T}} \succ 0$ and $R = R^{\mathrm{T}} \succ 0$. The optimal controller of the discrete-time LQR is

$$u^*(x_k) = -L^* x_k = -(R + B^{\mathrm{T}} V^* B)^{-1} B^{\mathrm{T}} V^* A x_k, \tag{3}$$

where $V^* = (V^*)^{\mathrm{T}} \succ 0$ is the unique solution to the following discrete-time algebraic Riccati equation (ARE)

$$A^{\mathrm{T}} V^* A - V^* - A^{\mathrm{T}} V^* B (R + B^{\mathrm{T}} V^* B)^{-1} B^{\mathrm{T}} V^* B \\ + Q = 0. \tag{4}$$

For a stabilizing control gain $L \in \mathbb{R}^{m\times n}$, the corresponding cost in (2) is $J_d(x_0, -Lx) = x_0^{\mathrm{T}} V_L x_0$, where

$V_L = (V_L)^{\mathrm{T}} \succ 0$ is the unique solution of the following Lyapunov equation

$$\mathcal{H}(G(V_L), L) = 0, \tag{5}$$

and the function $G(\cdot) : \mathbb{S}^n \to \mathbb{S}^{n+m}$ is defined as

$$\begin{aligned} G(V_L) &:= \begin{bmatrix} Q + A^{\mathrm{T}} V_L A & A^{\mathrm{T}} V_L B \\ B^{\mathrm{T}} V_L A & R + B^{\mathrm{T}} V_L B \end{bmatrix} \\ &= \begin{bmatrix} G_{xx}(V_L) & G_{ux}^{\mathrm{T}}(V_L) \\ G_{ux}(V_L) & G_{uu}(V_L) \end{bmatrix}. \end{aligned} \tag{6}$$

The discrete-time PI algorithm was developed by [7] to iteratively solve the discrete-time LQR problem. Given an initial stabilizing control gain L_0, the discrete-time PI algorithm is represented as:

Procedure 1 (Exact PI for discrete-time LQR)

1. *Policy evaluation: get* $G(V_i)$ *by solving*

$$\mathcal{H}(G(V_i), L_i) = 0. \tag{7}$$

2. *Policy improvement: get the improved policy by*

$$L_{i+1} = G_{uu}^{-1}(V_i) G_{ux}(V_i). \tag{8}$$

The monotonic convergence property of the discrete-time PI is shown in the following lemma.

Lemma 1 [7] *Given an initial stabilizing control gain* $L_0 \in \mathbb{R}^{m \times n}$*, the sequences* $\{V_i\}_{i=0}^{\infty}$ *and* $\{L_i\}_{i=0}^{\infty}$ *generated by iteratively solving* (7) *and* (8) *satisfy:*

1. $A - BL_i$ *is Schur for any* $i \in \mathbb{Z}_+$;
2. $V^* \preceq V_{i+1} \preceq V_i$;
3. $\lim_{i\to\infty} \left\| L_i - L^* \right\|_F = 0$ *and* $\lim_{i\to\infty} \left\| V_i - V^* \right\|_F = 0$.

2.2 Policy iteration for continuous-time LQR

Consider the continuous-time LTI system

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \tag{9}$$

where $x(t) \in \mathbb{R}^n$ is the state; $u(t) \in \mathbb{R}^m$ is the control input; x_0 is the initial state; A and B are constant matrices with compatible dimensions. The cost of system (9) is

$$J_c(x_0, u) = \int_0^{\infty} x^{\mathrm{T}}(t) Q x(t) + u^{\mathrm{T}}(t) R u(t) \mathrm{d}t. \tag{10}$$

Under Assumption 1, the classical continuous-time LQR aims at computing the optimal control policy as a function of the current state such that $J_c(x_0, u)$ is minimized. The optimal control policy is

$$u^*(x(t)) = -K^* x(t) = -R^{-1} B^{\mathrm{T}} P^* x(t), \tag{11}$$

where $P^* = (P^*)^{\mathrm{T}} \succ 0$ is the unique solution of the continuous-time ARE [26]:

$$A^{\mathrm{T}} P^* + P^* A - P^* B R^{-1} B^{\mathrm{T}} P^* + Q = 0. \tag{12}$$

For a stabilizing control gain $K \in \mathbb{R}^{m \times n}$, the corresponding cost in (10) is $J_c(x_0, -Kx) = x_0^{\mathrm{T}} P_K x_0$, where $P_K = (P_K)^{\mathrm{T}} \succ 0$ is the unique solution of the following Lyapunov equation

$$\mathcal{H}(M(P_K), K) = 0, \tag{13}$$

and the function $M(\cdot) : \mathbb{S}^n \to \mathbb{S}^{n+m}$ is defined as

$$\begin{aligned} M(P_K) &:= \begin{bmatrix} Q + A^{\mathrm{T}} P_K + P_K A & P_K B \\ B^{\mathrm{T}} P_K & R \end{bmatrix} \\ &= \begin{bmatrix} M_{xx}(P_K) & M_{ux}^{\mathrm{T}}(P_K) \\ M_{ux}(P_K) & M_{uu}(P_K) \end{bmatrix}. \end{aligned} \tag{14}$$

Given an initial stabilizing control gain K_0, the celebrated continuous-time PI developed in [6] iteratively solves the continuous-time LQR problem. The continuous-time PI algorithm is represented as:

Procedure 2 (Exact PI for continuous-time LQR)

1. *Policy evaluation: get* $M(P_i)$ *by solving*

$$\mathcal{H}(M(P_i), K_i) = 0. \tag{15}$$

2. *Policy improvement: get the improved policy by*

$$K_{i+1} = M_{uu}^{-1}(P_i) M_{ux}^{-1}(P_i). \tag{16}$$

Given an initial stabilizing control gain K_0, by iteratively solving (15) and (16), P_i monotonically converges to P^* and $(A - BK_i)$ is Hurwitz, which is formally presented in the following lemma.

Lemma 2 [6] *Given an initial stabilizing control gain* $K_0 \in \mathbb{R}^{m \times n}$*, the sequences* $\{K_i\}_{i=0}^{\infty}$ *and* $\{P_i\}_{i=0}^{\infty}$ *generated by iteratively solving* (15) *and* (16) *satisfy:*

1. $A - BK_i$ *is Hurwitz for any* $i \in \mathbb{Z}_+$;
2. $P^* \preceq P_{i+1} \preceq P_i$;
3. $\lim_{i\to\infty} \left\| K_i - K^* \right\|_F = 0$ *and* $\lim_{i\to\infty} \left\| P_i - P^* \right\|_F = 0$.

2.3 Problem formulation

For the discrete-time and continuous-time PI algorithms, the accurate model knowledge (A, B) is required for the algorithm implementation. The convergence of the PI algorithms in Lemmas 1 and 2 are based on the assumption that the accurate system model is attainable. However, in reality, system uncertainties are unavoidable, and the PI algorithms cannot be implemented exactly. Therefore, in this paper, we investigate the following problem.

Problem 1 *When the policy evaluation and improvement steps of the PI algorithms are subject to noise, will the convergence properties in Lemmas 1 and 2 still hold?*

3 Robustness analysis of policy iteration

In this section, we will formally introduce the inexact PI algorithms for the discrete-time and continuous-time LQR in the presence of noise. By invoking the concept of input-to-state stability [23], it is rigorously shown that the optimized control policies converge to a neighborhood of the optimal control policy, and the size of the neighborhood depends on the magnitude of the noise.

3.1 Robustness analysis of discrete-time policy iteration

According to the exact discrete-time PI algorithm in (7) and (8), in the presence of noise, the steps of policy evaluation and policy improvement cannot be implemented accurately, and the inexact PI algorithm is as follows.

Procedure 3 (Inexact PI for discrete-time LQR)

1. *Inexact policy evaluation: get $\hat{G}_i \in \mathbb{S}^{m+n}$ as an approximation of $G(\hat{V}_i)$, where $\hat{V}_i$ is the solution of*

$$\mathcal{H}(G(\hat{V}_i), \hat{L}_i) = 0. \tag{17}$$

2. *Inexact policy improvement: get the improved policy by*

$$\hat{L}_{i+1} = \hat{G}_{uu,i}^{-1}\hat{G}_{ux,i}. \tag{18}$$

In Procedure 3, $\Delta G_i = \hat{G}_i - G(\hat{V}_i)$ denotes the noise causing the inexact implementation of the PI algorithm. The "hat" is used to distinguish the cost matrices and control gains of the inexact PI from the exact PI. If $\Delta G_i = 0$, the control gain will be updated to the desired value $\bar{L}_{i+1} = G_{uu,i}^{-1}(\hat{V}_i)G_{ux,i}(\hat{V}_i)$. The noise ΔG_i causes the deviation between the updated control gain $\hat{L}_{i+1}$ and the desired control gain $\bar{L}_{i+1}$, i.e.

$$\begin{aligned}\Delta L_{i+1} &= \hat{L}_{i+1} - \bar{L}_{i+1}\\ &= (G_{uu}(\hat{V}_i) + \Delta G_{uu,i})^{-1}(G_{ux}(\hat{V}_i)\\ &\quad + \Delta G_{ux,i}) - G_{uu}^{-1}(\hat{V}_i)G_{ux}(\hat{V}_i).\end{aligned} \tag{19}$$

Remark 1 The noise ΔG_i can be caused by various factors. For example, in data-driven control [24], the matrix $G(\hat{V}_i)$ is identified by the collected input-state data. Since noise possibly pollutes the collected data, $\hat{G}_i$, instead of $G(\hat{V}_i)$, is obtained. Other factors that may cause ΔG_i include the inaccurate system identification, the residual error of numerically solving the Lyapunov equation, and the approximate values of Q and R in inverse optimal control in the absence of the exact knowledge of the cost function.

Next, by considering the inexact PI as a nonlinear dynamical system with the state $\hat{V}_i$, we analyze its robustness to noise ΔG_i by Lyapunov's direct method and in the sense of small-disturbance ISS. For any stabilizing control gain L, define the candidate Lyapunov function as

$$\mathcal{V}_d(V_L) = \mathrm{Tr}\,(V_L) - \mathrm{Tr}\left(V^*\right), \tag{20}$$

where $V_L = V_L^{\mathrm{T}} \succ 0$ is the solution of (5). Since $V_L \succeq V^*$ (obtained by Lemma 1), we have

$$\left\|V_L - V^*\right\|_F \le \mathcal{V}_d(V_L) \le \sqrt{n}\left\|V_L - V^*\right\|_F. \tag{21}$$

Remark 2 Since $J_d(x_0, -Lx) = x_0^{\mathrm{T}} V_L x_0$, when $x_0 \sim \mathcal{N}(0, I_n)$, $\mathbb{E}_{x_0} J_d(x_0, -Lx) = \mathrm{Tr}\,(V_L)$. Hence, the candidate Lyapunov function in (20) can be considered as the difference between the value function of the controller $u(x(t)) = -Lx(t)$ and the optimal value function.

For any $h > 0$, define a sublevel set $\mathcal{L}_h = \{L \in \mathbb{R}^{m\times n} | (A - BL) \text{ is Schur}, \mathcal{V}_d(V_L) \le h\}$. Since V_L is continuous with respect to the stabilizing control gain L, it readily follows that $\mathcal{L}_h$ is compact. Before the main theorem about the robustness of Procedure 3, we introduce the following instrumental lemma, which provides an upperbound on $\mathcal{V}_d(V_L)$.

Lemma 3 *For any stabilizing control gain L, let $L' = (R + B^{\mathrm{T}}V_LB)^{-1}B^{\mathrm{T}}V_LA$ and $E_L = (L'-L)^{\mathrm{T}}(R+B^{\mathrm{T}}V_LB)(L'-L)$. Then,*

$$\mathcal{V}_d(V_L) \le a\left\|E_L\right\|, \tag{22}$$

where

$$a = \mathrm{Tr}\left(\sum_{k=0}^{\infty}(A - BL^*)^{k,\mathrm{T}}(A - BL^*)^k\right). \tag{23}$$

Proof We can rewrite (4) as

$$(A - BL^*)^{\mathrm{T}} V^* (A - BL^*) - V^* + Q + L^{*T} R L^* = 0. \tag{24}$$

In addition, it follows from (5) that

$$\begin{aligned}&(A - BL^*)^{\mathrm{T}} V_L (A - BL^*) - V_L + Q + E_L + L^{*T} R L^* \\ &\quad - (L' - L^*)^{\mathrm{T}} (R + B^{\mathrm{T}} V_L B)(L' - L^*) = 0.\end{aligned} \tag{25}$$

Subtracting (24) from (25) yields

$$\begin{aligned}&(A - BL^*)^{\mathrm{T}} (V_L - V^*)(A - BL^*) - (V_L - V^*) \\ &\quad + E_L - (L' - L^*)^{\mathrm{T}} (R + B^{\mathrm{T}} V_L B)(L' - L^*) = 0.\end{aligned} \tag{26}$$

Since $(A - BL^*)$ is Schur, it follows from [27, Theorem 5.D6] that

$$V_L - V^* \preceq \sum_{k=0}^{\infty} (A - BL^*)^{k,\mathrm{T}} E_L (A - BL^*)^k. \tag{27}$$

Taking the trace of (27) and using the main result of [28], we have

$$\mathcal{V}_d(V_L) \le \left\| E_L \right\| \mathrm{Tr} \left(\sum_{k=0}^{\infty} (A - BL^*)^{k,\mathrm{T}} (A - BL^*)^k \right). \tag{28}$$

Hence, the proof is completed. □

Lemma 4 *For any $L \in \mathcal{L}_h$,*

$$\mathrm{Tr} \left(\sum_{k=0}^{\infty} (A - BL)^{k,\mathrm{T}} (A - BL)^k \right) \le \frac{h + \mathrm{Tr}\,(V^*)}{\underline{\lambda}(Q)}. \tag{29}$$

Proof Since $(A - BL)$ is Schur, it follows from (5) and [27, Theorem 5.D6] that

$$V_L = \sum_{k=0}^{\infty} (A - BL)^{k,\mathrm{T}} (Q + L^{\mathrm{T}} R L)(A - BL)^k. \tag{30}$$

Taking the trace of (30), and using the result of [28], we have

$$h + \mathrm{Tr}\left(V^*\right) \ge \underline{\lambda}(Q) \mathrm{Tr} \left(\sum_{k=0}^{\infty} (A - BL)^{k,\mathrm{T}} (A - BL)^k \right). \tag{31}$$

Hence, (29) readily follows from (31). □

The following lemma shows that the sublevel set $\mathcal{L}_h$ is invariant as long as the noise ΔL_i is sufficiently small.

Lemma 5 *For any $h > 0$ and given an initial control gain $\hat{L}_0 \in \mathcal{L}_h$, if $\left\| \Delta L \right\|_{\infty} < b(h)$, where $b(h)$ is defined in (44), then, $\hat{L}_i \in \mathcal{L}_h$ for any $i \in \mathbb{Z}_+$.*

Proof Induction is applied to prove the statement. When $i = 0$, $\hat{L}_0 \in \mathcal{L}_h$. Suppose $\hat{L}_i \in \mathcal{L}_h$, then, by [27, Theorem 5.D6], we have $\hat{V}_i \succ 0$. We can rewrite (17) as

$$\begin{aligned}&(A - B\hat{L}_{i+1})^{\mathrm{T}} \hat{V}_i (A - B\hat{L}_{i+1}) - \hat{V}_i + Q + \hat{L}_i^{\mathrm{T}} R \hat{L}_i \\ &\quad + (\hat{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i (A - B\hat{L}_{i+1}) \\ &\quad + (A - B\hat{L}_{i+1})^{\mathrm{T}} \hat{V}_i B (\hat{L}_{i+1} - \hat{L}_i) \\ &\quad + (\hat{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i B (\hat{L}_{i+1} - \hat{L}_i) = 0.\end{aligned} \tag{32}$$

Considering (19), we have

$$\begin{aligned}&(\hat{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i (A - B\hat{L}_{i+1}) \\ &\quad = (\bar{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} R \bar{L}_{i+1} + \Delta L_{i+1}^{\mathrm{T}} R \bar{L}_{i+1} \\ &\qquad - \Delta L_{i+1}^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i B \Delta L_{i+1} - (\bar{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i B \Delta L_{i+1}.\end{aligned} \tag{33}$$

In addition, it follows from (19) that

$$\begin{aligned}&(\hat{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i B (\hat{L}_{i+1} - \hat{L}_i) \\ &\quad = (\bar{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i B (\bar{L}_{i+1} - \hat{L}_i) \\ &\qquad + \Delta L_{i+1}^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i B \Delta L_{i+1} + \Delta L_{i+1}^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i B (\bar{L}_{i+1} - \hat{L}_i) \\ &\qquad + (\bar{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} B^{\mathrm{T}} \hat{V}_i B \Delta L_{i+1}.\end{aligned} \tag{34}$$

Plugging (33) and (34) into (32), and completing the squares, we have

$$\begin{aligned}&(A - B\hat{L}_{i+1})^{\mathrm{T}} \hat{V}_i (A - B\hat{L}_{i+1}) - \hat{V}_i + Q + \hat{L}_{i+1}^{\mathrm{T}} R \hat{L}_{i+1} \\ &\quad + \hat{E}_i - \Delta L_{i+1}^{\mathrm{T}} (R + B^{\mathrm{T}} \hat{V}_i B) \Delta L_{i+1} = 0,\end{aligned} \tag{35}$$

where

$$\hat{E}_i = (\bar{L}_{i+1} - \hat{L}_i)^{\mathrm{T}} (R + B^{\mathrm{T}} \hat{V}_i B)(\bar{L}_{i+1} - \hat{L}_i). \tag{36}$$

If

$$\left\| \Delta L_{i+1} \right\| < \sqrt{\frac{\underline{\lambda}(Q)}{\left\| R \right\| + \left\| B \right\|^2 (h + \mathrm{Tr}\,(V^*))}} =: b_1(h), \tag{37}$$

it is guaranteed that

$$Q - \Delta L_{i+1}^{\mathrm{T}} (R + B^{\mathrm{T}} \hat{V}_i B) \Delta L_{i+1} \succ 0. \tag{38}$$

Hence, if $\left\| \Delta L_{i+1} \right\| < b_1(h)$, it follows from (35) and [29, Theorem 8.4] that $A - B\hat{L}_{i+1}$ is Schur.

Writing down (17) at the $(i+1)$th iteration, and subtracting it from (35), we have

$$(A - B\hat{L}_{i+1})^{\mathrm{T}}(\hat{V}_i - \hat{V}_{i+1})(A - B\hat{L}_{i+1}) - (\hat{V}_i - \hat{V}_{i+1}) + \hat{E}_i - \Delta L_{i+1}^{\mathrm{T}}(R + B^{\mathrm{T}}\hat{V}_i B)\Delta L_{i+1} = 0. \tag{39}$$

Following [27, Theorem 5.D6], we have

$$\begin{aligned}&\hat{V}_{i+1} - V^* \\ &\quad = \hat{V}_i - V^* - \sum_{k=0}^{\infty}(A - B\hat{L}_{i+1})^{k,\mathrm{T}}\hat{E}_i(A - B\hat{L}_{i+1})^k \\ &\qquad + \sum_{k=0}^{\infty}(A - B\hat{L}_{i+1})^{k,\mathrm{T}}\Delta L_{i+1}^{\mathrm{T}}(R \\ &\qquad + B^{\mathrm{T}}\hat{V}_i B)\Delta L_{i+1}(A - B\hat{L}_{i+1})^k.\end{aligned} \tag{40}$$

Taking the trace of (40), and using Lemma 3 and the result in [28] yield

$$\begin{aligned}&\mathcal{V}_d(\hat{V}_{i+1}) \\ &\quad \leq (1 - \frac{1}{a})\mathcal{V}_d(\hat{V}_i) + \mathrm{Tr}\Big(\sum_{k=0}^{\infty}(A - B\hat{L}_{i+1})^{k,\mathrm{T}}(R \\ &\qquad + B^{\mathrm{T}}\hat{V}_i B)(A - B\hat{L}_{i+1})^k\Big)\|\Delta L_{i+1}\|^2.\end{aligned} \tag{41}$$

It follows from (41), Lemma 4 and [28] that

$$\mathcal{V}_d(\hat{V}_{i+1}) \leq (1 - \frac{1}{a})\mathcal{V}_d(\hat{V}_i) + b_2(h)\|\Delta L_{i+1}\|^2, \tag{42}$$

where b_2 is defined as

$$b_2(h) = \frac{(h + \mathrm{Tr}\,(V^*))\|R\| + \|B\|^2(h + \mathrm{Tr}\,(V^*))^2}{\underline{\lambda}(Q)}. \tag{43}$$

Hence, if

$$\|\Delta L_{i+1}\| < \sqrt{\frac{h}{ab_2(h)}} =: b(h), \tag{44}$$

it is guaranteed that

$$\mathcal{V}_d(\hat{V}_{i+1}) \leq h. \tag{45}$$

In addition, it is observed that $b(h) = \sqrt{\frac{h}{h+\mathrm{Tr}(V^*)}}\sqrt{\frac{1}{a}} \times b_1(h) \leq b_1(h)$.

In summary, if $\|\Delta L_{i+1}\| < b(h)$, it follows from (37) and (44) that $\hat{L}_{i+1} \in \mathcal{L}_h$. The lemma is consequently proved by induction. □

Now, by Lypaunov's direct method and by viewing Procedure 3 as a discrete-time nonlinear system with the state $\hat{V}_i$, it is shown that $\hat{V}_i$ converges to a small neighbourhood of the optimal solution as long as noise is sufficiently small.

Lemma 6 *For any $h > 0$ and $\hat{L}_0 \in \mathcal{L}_h$, if $\|\Delta L\|_\infty < b(h)$, there exist a $\mathcal{K}$-function $\rho(\cdot)$ and a $\mathcal{KL}$-function $\kappa(\cdot,\cdot)$, such that*

$$\|\hat{V}_i - V^*\|_F \leq \kappa(\|\hat{V}_0 - V^*\|_F, i) + \rho(\|\Delta L\|_\infty). \tag{46}$$

Proof Repeating (42) for $i, i-1, \ldots, 0$, we have

$$\mathcal{V}_d(\hat{V}_i) \leq \left(1 - \frac{1}{a}\right)^i \mathcal{V}_d(\hat{V}_0) + ab_2(h)\|\Delta L\|_\infty^2. \tag{47}$$

Considering (21), it follows that

$$\|\hat{V}_i - V^*\|_F \leq \sqrt{n}\left(1 - \frac{1}{a}\right)^i \|\hat{V}_0 - V^*\|_F + ab_2(h)\|\Delta L\|_\infty^2. \tag{48}$$

The proof is thus completed. □

The small-disturbance ISS property of the Procedure 3 is shown in the following theorem.

Theorem 1 *The inexact PI in Procedure 3 is small-disturbance ISS. That is, for any $h > 0$ and $\hat{L}_0 \in \mathcal{L}_h$, if $\|\Delta G\|_\infty < \min\{\frac{b_3}{2}, \frac{b(h)}{b_5}\}$, where b_3 and $b_5(h)$ are defined in (50) and (53) respectively, then,*

$$\|\hat{V}_i - V^*\|_F \leq \kappa(\|\hat{V}_0 - V^*\|_F, i) + \rho(b_5(h)\|\Delta G\|_\infty). \tag{49}$$

Proof Suppose $\hat{L}_i \in \mathcal{L}_h$. Since $\hat{V}_i \succeq V^*$, we have $G_{uu}(\hat{V}_i) \succeq R + B^{\mathrm{T}}V^*B$. Therefore, if

$$\|\Delta G_{uu,i}\| < \underline{\lambda}(R + B^{\mathrm{T}}V^*B) =: b_3, \tag{50}$$

$(\Delta G_{uu,i} + G_{uu}(\hat{V}_i))$ is invertible. It follows from (19) and

$$\begin{aligned}&(G_{uu}(\hat{V}_i) + \Delta G_{uu,i})^{-1} \\ &\quad = G_{uu}^{-1}(\hat{V}_i) - G_{uu}^{-1}(\hat{V}_i)\Delta G_{uu,i}(G_{uu}(\hat{V}_i) + \Delta G_{uu,i})^{-1}\end{aligned}$$

that

$$\begin{aligned}&\|\Delta L_{i+1}\| \\ &\quad \leq \|G_{uu}^{-1}(\hat{V}_i)\|\Big(\|\Delta G_{ux,i}\| + \|(G_{uu}(\hat{V}_i) + \Delta G_{uu,i})^{-1}\| \\ &\qquad \times \|G_{ux}(\hat{V}_i) + \Delta G_{ux,i}\|\|\Delta G_{uu,i}\|\Big).\end{aligned} \tag{51}$$

If $\|\Delta G_i\| < b_3/2$, we have

$$\begin{aligned}
&\|G_{uu}^{-1}(\hat{V}_i)\| < \frac{1}{b_3},\\
&\|(G_{uu}(\hat{V}_i) + \Delta G_{uu,i})^{-1}\| < \frac{2}{b_3},\\
&\|G_{ux}(\hat{V}_i) + \Delta G_{ux,i}\| < \|A\| \|B\|(h + \mathrm{Tr}\,(V^*)) + b_3/2\\
&\qquad\qquad =: b_4(h).
\end{aligned} \tag{52}$$

Consequently,

$$\|\Delta L_{i+1}\| < \frac{b_3 + 2b_4(h)}{b_3^2}\|\Delta G_i\| =: b_5(h)\|\Delta G_i\|. \tag{53}$$

Therefore, if $\|\Delta G_i\| < \frac{b(h)}{b_5(h)}$, it is guaranteed that $\|\Delta L_{i+1}\| < b(h)$. Following (44) and (45), we have $\hat{L}_{i+1} \in \mathcal{L}_h$. Since $\hat{L}_0 \in \mathcal{L}_h$, we have $\hat{L}_i \in \mathcal{L}_h$ for any $i \in \mathbb{Z}_+$, which implies that $\|\Delta L\|_\infty < b(h)$ given that $\|\Delta G\|_\infty < \frac{b(h)}{b_5(h)}$.

It follows from (53) that $\|\Delta L\|_\infty < b_5(h)\|\Delta G\|_\infty$. By Lemma 6, the proof is thus completed. □

3.2 Robustness analysis of continuous-time policy iteration

According to the exact PI for continuous-time LQR in (15) and (16), in the presence of noise, the steps of policy evaluation and policy improvement cannot be implemented accurately, and the inexact PI is as follows.

Procedure 4 (Inexact PI for continuous-time LQR)

1. *Inexact policy evaluation: get $\hat{M}_i \in \mathbb{S}^{m+n}$ as an approximation of $M(\hat{P}_i)$, where $\hat{P}_i$ is the solution of*

$$\mathcal{H}(M(\hat{P}_i), \hat{K}_i) = 0. \tag{54}$$

2. *Inexact policy improvement: get the updated control gain by*

$$\hat{K}_{i+1} = \hat{M}_{uu,i}^{-1}\hat{M}_{ux,i}. \tag{55}$$

In Procedure 4, $\Delta M_i = \hat{M}_i - M(V_i)$ denotes the noise causing the inexact implementation of the PI algorithm. The "hat" is used to distinguish the cost matrices and control gains of the inexact PI from the exact PI. If $\Delta M_i = 0$, the control gain will be updated to the desired value $\bar{K}_{i+1} = M_{uu,i}^{-1}(\hat{P}_i)M_{ux,i}(\hat{P}_i)$. The noise ΔM_i causes the deviation between the updated control gain $\hat{K}_{i+1}$ and the desired control gain $\bar{K}_{i+1}$, i.e.

$$\begin{aligned}
&\Delta K_{i+1}\\
&\quad = \hat{K}_{i+1} - \bar{K}_{i+1}\\
&\quad = (M_{uu}(\hat{P}_i) + \Delta M_{uu,i})^{-1}(M_{ux}(\hat{P}_i) + \Delta M_{ux,i})\\
&\qquad - M_{uu}^{-1}(\hat{P}_i)M_{ux}(\hat{P}_i).
\end{aligned} \tag{56}$$

For any stabilizing control gain K, define the candidate Lyapunov function as

$$\mathcal{V}_c(P_K) = \mathrm{Tr}\,(P_K) - \mathrm{Tr}\,\left(P^*\right), \tag{57}$$

where $P_K = P_K^{\mathrm{T}} \succ 0$ is the solution of (13), i.e.

$$\begin{aligned}
&(A - BK)^{\mathrm{T}}P_K + P_K(A - BK) + Q + K^{\mathrm{T}}RK\\
&\quad = 0.
\end{aligned} \tag{58}$$

Since $P_K \succeq P^*$, we have

$$\|P_K - P^*\|_F \leq \mathcal{V}_c(P_K) \leq \sqrt{n}\|P_K - P^*\|_F. \tag{59}$$

For any $h > 0$, define the sublevel set $\mathcal{K}_h = \{K \in \mathbb{R}^{m\times n} | (A - BK)$ is Hurwitz, $\mathcal{V}_c(P_K) \leq h\}$. Since P_K is continuous with respect to the stabilizing control gain K, the sublevel set $\mathcal{K}_h$ is compact.

The following lemmas are instrumental for the proof of the main theorem.

Lemma 7 *Consider a Hurwitz matrix $D \in \mathbb{R}^{n\times n}$ and a positive semi-definite matrix $E \in \mathbb{S}^n$. Define $H(D, E) = \int_0^\infty e^{D^{\mathrm{T}}t}Ee^{Dt}\mathrm{d}t$, and $c(D) = \log(5/4)/\|D\|$. Then,*

$$\|H(D, E)\| \geq \frac{1}{2}c(D)\|E\|. \tag{60}$$

Proof The Taylor expansion of e^{Dt} is

$$e^{Dt} = I_n + \sum_{k=1}^{\infty}(Dt)^k/k! = I_n + F(t). \tag{61}$$

Hence,

$$\|F(t)\| \leq \sum_{k=1}^{\infty}(\|D\|t)^k/k! = e^{\|D\|t} - 1. \tag{62}$$

Pick a $v \in \mathbb{R}^n$ which satisfies $v^{\mathrm{T}} E v = \|E\||v|^2$. Then,

$$\begin{aligned} v^{\mathrm{T}} H v &\geq \int_0^{c(D)} v^{\mathrm{T}} e^{D^{\mathrm{T}} t} E e^{Dt} v \mathrm{d}t \\ &= \int_0^{c(D)} v^{\mathrm{T}} (I_n + F(t))^{\mathrm{T}} E (I_n + F(t)) v \mathrm{d}t \\ &\geq \int_0^{c(D)} \|E\||v|^2 - 2\|F(t)\|\|E\||v|^2 \mathrm{d}t \\ &\geq \int_0^{c(D)} \|E\||v|^2 - 2(e^{\|D\|t} - 1)\|E\||v|^2 \mathrm{d}t \\ &\geq \int_0^{c(D)} (3 - 2e^{\|D\|c(D)})\|E\||v|^2 \mathrm{d}t \\ &= \frac{1}{2} c(D)\|E\||v|^2. \end{aligned} \tag{63}$$

Hence, the lemma follows readily from (63). □

The following lemma presents an upperbound of the Lyapunov function $\mathcal{V}_c(P_K)$.

Lemma 8 *For any stabilizing control gain K, let $K' = R^{-1} B^{\mathrm{T}} P_K$, where $P_K = P_K^{\mathrm{T}} \succ 0$ is the solution of (58), and $E_K = (K' - K)^{\mathrm{T}} R (K' - K)$. Then,*

$$\begin{aligned} &\mathcal{V}_c(P_K) \leq d\|E_K\|, \\ &d = \mathrm{Tr}\left(\int_0^{\infty} e^{(A-BK^*)^{\mathrm{T}} t} e^{(A-BK^*)t} \mathrm{d}t\right). \end{aligned} \tag{64}$$

Proof Rewrite ARE (12) as

$$\begin{aligned} &(A - BK^*)^{\mathrm{T}} P^* + P^*(A - BK^*) + Q + (K^*)^{\mathrm{T}} R K^* \\ &\quad = 0. \end{aligned} \tag{65}$$

Furthermore, (58) is rewritten as

$$\begin{aligned} &(A - BK^*)^{\mathrm{T}} P_K + P_K (A - BK^*) + Q + K^{\mathrm{T}} R K \\ &\quad + (K^* - K)^{\mathrm{T}} B^{\mathrm{T}} P_K + P_K B (K^* - K) = 0. \end{aligned} \tag{66}$$

Subtracting (65) from (66) yields

$$\begin{aligned} &(A - BK^*)^{\mathrm{T}} (P_K - P^*) + (P_K - P^*)(A - BK^*) \\ &\quad + K^{\mathrm{T}} R K - (K^*)^{\mathrm{T}} R K^* + (K^* - K)^{\mathrm{T}} B^{\mathrm{T}} P_K \\ &\quad + P_K B (K^* - K) = 0. \end{aligned} \tag{67}$$

Considering $K' = R^{-1} B^{\mathrm{T}} P_K$ and completing the squares in (67), we have

$$\begin{aligned} &(A - BK^*)^{\mathrm{T}} (P_K - P^*) + (P_K - P^*)(A - BK^*) + E_K \\ &\quad - (K' - K^*)^{\mathrm{T}} R (K' - K^*) = 0. \end{aligned} \tag{68}$$

Since $(A - BK^*)$ is Hurwitz, by (68) and [27, equation (5.18)], we have

$$P_K - P^* \preceq \int_0^{\infty} e^{(A-BK^*)^{\mathrm{T}} t} E_K e^{(A-BK^*)t} \mathrm{d}t. \tag{69}$$

Taking the trace of (69), considering the cyclic property of trace and [28], we have (64). □

Lemma 9 *For any $K \in \mathcal{K}_h$,*

$$\mathrm{Tr}\left(\int_0^{\infty} e^{(A-BK)^{\mathrm{T}} t} e^{(A-BK)t} \mathrm{d}t\right) \leq \frac{h + \mathrm{Tr}\,(P^*)}{\underline{\lambda}(Q)}. \tag{70}$$

Proof Since $A - BK$ is Hurwitz, it follows from (58) that

$$P_K = \int_0^{\infty} e^{(A-BK)^{\mathrm{T}} t} (Q + K^{\mathrm{T}} R K) e^{(A-BK)t} \mathrm{d}t. \tag{71}$$

Taking the trace of (71), and considering [28], we have

$$h + \mathrm{Tr}\,(P^*) \geq \mathrm{Tr}\left(\int_0^{\infty} e^{(A-BK)^{\mathrm{T}} t} e^{(A-BK)t} \mathrm{d}t\right) \underline{\lambda}(Q). \tag{72}$$

The proof is hence completed. □

The following lemma implies that the sublevel set $\mathcal{K}_h$ is invariant under the inexact PI in Procedure 4 as long as $\|\Delta K\|_{\infty}$ is sufficiently small.

Lemma 10 *For any $h > 0$ and given an initial control gain $\hat{K}_0 \in \mathcal{K}_h$, if $\|\Delta K\|_{\infty} < e(h)$, where $e(h)$ is defined in (83), then, $\hat{K}_i \in \mathcal{K}_h$ for any $i \in \mathbb{Z}_+$.*

Proof Induction is used to prove the statement. When $i = 0$, $\hat{K}_0 \in \mathcal{K}_h$. Suppose that $\hat{K}_i \in \mathcal{K}_h$, then, by [30, Lemma 3.18], $\hat{P}_i \succ 0$. We can rewrite (54) as

$$\begin{aligned} &(A - B\hat{K}_{i+1})^{\mathrm{T}} \hat{P}_i + \hat{P}_i (A - B\hat{K}_{i+1}) + Q + \hat{K}_i^{\mathrm{T}} R \hat{K}_i \\ &(\hat{K}_{i+1} - \hat{K}_i)^{\mathrm{T}} B^{\mathrm{T}} \hat{P}_i + \hat{P}_i B (\hat{K}_{i+1} - \hat{K}_i) = 0. \end{aligned} \tag{73}$$

Considering (56), we have

$$\begin{aligned} &(A - B\hat{K}_{i+1})^{\mathrm{T}} \hat{P}_i + \hat{P}_i (A - B\hat{K}_{i+1}) + Q + \hat{K}_i^{\mathrm{T}} R \hat{K}_i \\ &\quad + (\bar{K}_{i+1} - \hat{K}_i)^{\mathrm{T}} R \bar{K}_{i+1} + \bar{K}_{i+1}^{\mathrm{T}} R (\bar{K}_{i+1} - \hat{K}_i) \\ &\quad + \Delta K_{i+1}^{\mathrm{T}} R \bar{K}_{i+1} + \bar{K}_{i+1}^{\mathrm{T}} R \Delta K_{i+1} = 0. \end{aligned} \tag{74}$$

Completing the squares in (74) yields

$$\begin{aligned} &(A - B\hat{K}_{i+1})^{\mathrm{T}} \hat{P}_i + \hat{P}_i (A - B\hat{K}_{i+1}) + Q \\ &\quad + \hat{K}_{i+1}^{\mathrm{T}} R \hat{K}_{i+1} + \hat{E}_i - \Delta K_{i+1}^{\mathrm{T}} R \Delta K_{i+1} = 0, \end{aligned} \tag{75}$$

where

$$\hat{E}_i = (\bar{K}_{i+1} - \hat{K}_i)^{\mathrm{T}} R (\bar{K}_{i+1} - \hat{K}_i). \tag{76}$$

Hence, by [30, Lemma 3.19], $A - B\hat{K}_{i+1}$ is Hurwitz as long as $\|\Delta K_i\| \leq e_1$, where e_1 is defined as

$$e_1 = \sqrt{\underline{\lambda}(Q)/\|R\|}. \tag{77}$$

Writing down (54) for the $(i+1)$th iteration, and subtracting it from (75), we have

$$\begin{aligned}&(A - B\hat{K}_{i+1})^{\mathrm{T}}(\hat{P}_i - \hat{P}_{i+1}) + (\hat{P}_i - \hat{P}_{i+1})(A - B\hat{K}_{i+1}) \\ &\quad + \hat{E}_i - \Delta K_{i+1}^{\mathrm{T}} R \Delta K_{i+1} = 0.\end{aligned} \tag{78}$$

Since $(A - B\hat{K}_{i+1})$ is Hurwitz ($e(h) \leq e_1(h)$), it follows from [27, equation (5.18)] and (78) that

$$\begin{aligned}&\hat{P}_i - \hat{P}_{i+1} \\ &\quad = \int_0^\infty e^{(A-B\hat{K}_{i+1})^{T,t}} (\hat{E}_i - \Delta K_{i+1}^{\mathrm{T}} R \Delta K_{i+1}) e^{(A-B\hat{K}_{i+1})} \mathrm{d}t.\end{aligned} \tag{79}$$

Taking the trace of (79), and considering [28] and Lemma 9, we have

$$\begin{aligned}\mathcal{V}_c(\hat{P}_{i+1}) \leq \mathcal{V}_c(\hat{P}_i) &- \|H(A - B\hat{K}_{i+1}, \hat{E}_i)\| \\ &+ \frac{h + \mathrm{Tr}\,(P^*)}{\underline{\lambda}(Q)} \|R\| \|\Delta K_{i+1}\|^2.\end{aligned} \tag{80}$$

It follows from (80) and Lemmas 7 and 8 that

$$\begin{aligned}\mathcal{V}_c(\hat{P}_{i+1}) \leq &\left(1 - \frac{\log{(5/4)}}{2d\|A - B\hat{K}_{i+1}\|}\right) \mathcal{V}_c(\hat{P}_i) \\ &+ \frac{h + \mathrm{Tr}\,(P^*)}{\underline{\lambda}(Q)} \|R\| \|\Delta K_{i+1}\|^2.\end{aligned} \tag{81}$$

Taking the expression of $\hat{K}_{i+1}$ into consideration, we have

$$\begin{aligned}&\|A - B\hat{K}_{i+1}\| \\ &\quad \leq \|A\| + \|BR^{-1}B^{\mathrm{T}}\|(\mathrm{Tr}\left(P^*\right) + h) + \|B\|e_1 \\ &\quad =: e_2(h).\end{aligned} \tag{82}$$

Plugging (82) into (81), it follows that if

$$\|\Delta K_{i+1}\| < \left(\frac{\log{(5/4)} h \underline{\lambda}(Q)}{2d\|R\|(h + \mathrm{Tr}\,(P^*))e_2(h)}\right)^{1/2} =: e(h), \tag{83}$$

we have

$$\mathcal{V}_c(\hat{P}_{i+1}) \leq h. \tag{84}$$

It is observed that $e(h) \leq e_1(h)$. Hence, Given $\|\Delta K\|_\infty < e(h)$, $\hat{K}_{i+1} \in \mathcal{K}_h$ and the proof is completed by induction. □

Lemma 11 *For any $h > 0$ and $\hat{K}_0 \in \mathcal{K}_h$, if $\|\Delta K\|_\infty < e(h)$, then, there exist a $\mathcal{KL}$-function $\alpha(\cdot, \cdot)$ and a $\mathcal{K}$-function $\beta(\cdot)$, such that*

$$\|\hat{P}_i - P^*\|_F \leq \alpha(\|\hat{P}_0 - P^*\|_F, i) + \beta(\|\Delta K\|_\infty). \tag{85}$$

Proof It follows from Lemma 10, (81) and (82) that for any $i \in \mathbb{Z}_+$,

$$\begin{aligned}\mathcal{V}_c(\hat{P}_i) \leq &\left(1 - \frac{\log{(5/4)}}{2de_2(h)}\right) \mathcal{V}_c(\hat{P}_{i-1}) \\ &+ \frac{h + \mathrm{Tr}\,(P^*)}{\underline{\lambda}(Q)} \|R\| \|\Delta K_i\|^2.\end{aligned} \tag{86}$$

Repeating (86) for $i, i-1, \ldots, 1, 0$, we have

$$\begin{aligned}\mathcal{V}_c(\hat{P}_i) \leq &\left(1 - \frac{\log{(5/4)}}{2de_2(h)}\right)^i \mathcal{V}_c(\hat{P}_0) \\ &+ \frac{h + \mathrm{Tr}\,(P^*)}{\underline{\lambda}(Q)} \frac{\|R\| 2de_2(h)}{\log{(5/4)}} \|\Delta K\|_\infty^2.\end{aligned} \tag{87}$$

By (59), we have

$$\begin{aligned}\|\hat{P}_i - P^*\|_F \leq &\left(1 - \frac{\log{(5/4)}}{2de_2(h)}\right)^i \sqrt{n}\|\hat{P}_0 - P^*\|_F \\ &+ \frac{h + \mathrm{Tr}\,(P^*)}{\underline{\lambda}(Q)} \frac{\|R\| 2de_2(h)}{\log{(5/4)}} \|\Delta K\|_\infty^2.\end{aligned} \tag{88}$$

Hence, (85) follows readily. □

With the aforementioned lemmas, we are ready to propose the main result on the robustness of the inexact PI for continuous-time LQR.

Theorem 2 *The inexact PI in Procedure 4 is small-disturbance ISS. That is, for any $h > 0$ and $\hat{K}_0 \in \mathcal{K}_h$, if $\|\Delta M\|_\infty < \min\{\frac{e_3}{2}, \frac{e(h)}{e_5(h)}\}$, where e_3 and $e_5(h)$ are defined in (90) and (93) respectively, then,*

$$\begin{aligned}&\|\hat{P}_i - P^*\|_F \\ &\quad \leq \alpha(\|\hat{P}_0 - P^*\|_F, i) + \beta(e_5(h)\|\Delta M\|_\infty).\end{aligned} \tag{89}$$

Proof Suppose $\hat{K}_i \in \mathcal{L}_h$. If

$$\|\Delta M_{uu,i}\| < \underline{\lambda}(R) =: e_3, \tag{90}$$

$(\Delta M_{uu,i} + M_{uu}(\hat{P}_i))$ is invertible. It follows from (56) and

$$(M_{uu}(\hat{P}_i) + \Delta M_{uu,i})^{-1}$$

$$= M_{uu}^{-1}(\hat{P}_i) - M_{uu}^{-1}(\hat{P}_i)\Delta M_{uu,i}(M_{uu}(\hat{P}_i) + \Delta M_{uu,i})^{-1}$$

that

$$\begin{aligned}\left\|\Delta K_{i+1}\right\| \leq &\left\|M_{uu}^{-1}(\hat{P}_i)\right\|\Big(\left\|\Delta M_{ux,i}\right\| + \big\|(M_{uu}(\hat{P}_i)\\ &+ \Delta M_{uu,i})^{-1}\big\|\left\|M_{ux}(\hat{P}_i)\right.\\ &\left.+ \Delta M_{ux,i}\right\|\left\|\Delta M_{uu,i}\right\|\Big).\end{aligned} \tag{91}$$

If $\left\|\Delta M_i\right\| < e_3/2$, we have

$$\begin{aligned}&\left\|M_{uu}^{-1}(\hat{P}_i)\right\| < \frac{1}{e_3},\\ &\left\|(M_{uu}(\hat{P}_i) + \Delta M_{uu,i})^{-1}\right\| < \frac{2}{e_3},\\ &\left\|M_{ux}(\hat{P}_i) + \Delta M_{ux,i}\right\| < \left\|B\right\|(h + \mathrm{Tr}\,(P^*)) + e_3/2\\ &\qquad\qquad =: e_4(h).\end{aligned} \tag{92}$$

Consequently,

$$\left\|\Delta K_{i+1}\right\| < \frac{e_3 + 2e_4(h)}{e_3^2}\left\|\Delta M_i\right\| =: e_5(h)\left\|\Delta M_i\right\|. \tag{93}$$

Therefore, if $\left\|\Delta M_i\right\| < \frac{e(h)}{e_5(h)}$, it is guaranteed that $\left\|\Delta K_{i+1}\right\| < e(h)$. Following (83) and (84), we have $\hat{K}_{i+1} \in \mathcal{K}_h$. Since $\hat{K}_0 \in \mathcal{K}_h$, we have $\hat{K}_i \in \mathcal{K}_h$ for any $i \in \mathbb{Z}_+$, which implies that $\left\|\Delta K\right\|_\infty < e(h)$ given that $\left\|\Delta M\right\|_\infty < \frac{e(h)}{e_5(h)}$.

It follows from (93) that $\left\|\Delta K\right\|_\infty < e_5(h)\left\|\Delta M\right\|_\infty$. By Lemma 11, the proof is thus completed. □

Remark 3 Compared with [24, 25], in Theorems 1 and 2, the explicit expressions of the upperbounds on the small disturbance are given, such that at each iteration, the generated control gain is stabilizing and contained in the sublevel sets $\mathcal{L}_h$ and $\mathcal{K}_h$. In addition, it is observed from (48) and (88) that the control gains generated by the inexact PI algorithms ultimately converge to a neighborhood of the optimal solution, and the size of the neighborhood is proportional to the quadratic form of the noise.

4 Learning-based policy iteration

In this section, based on the robustness property of the inexact PI in Procedure 3, we will develop a learning-based PI algorithm. Only the input-state trajectory data measured from the system is required for the algorithm.

4.1 Algorithm development

For a signal $u_{[0,N-1]} = [u_0, u_1, \ldots, u_{N-1}]$, its Hankel matrix of depth l is represented as

$$H_l(u_{[0,N-1]}) = \begin{bmatrix} u_0 & u_1 & \ldots & u_{N-l} \\ u_1 & u_2 & \ldots & u_{N-l+1} \\ \vdots & \vdots & & \vdots \\ u_{l-1} & u_l & \ldots & u_{N-1}\end{bmatrix}. \tag{94}$$

Definition 1 An input signal $u_{[0,N-1]}$ is persistent exciting (PE) of order l if the Hankel matrix $H_l(u_{[0,N-1]})$ is full row rank.

Lemma 12 [31] *Let an input signal $u_{[0,N-1]}$ be PE of order $l+n$. Then, the state trajectory $x_{[0,N-1]}$ sampled from system (1) driven by the input $u_{[0,N-1]}$ satisfies*

$$\mathrm{rank}\left(\begin{bmatrix} H_1(x_{[0,N-1]}) \\ H_l(u_{[0,N-1]})\end{bmatrix}\right) = lm + n. \tag{95}$$

Given the input-state data $u_{[0,N-1]}$ and $x_{[0,N]}$ sampled from (1), we will design a learning-based PI algorithm such that the accurate knowledge of system matrices is not required. For any time indices $0 \leq k_1, k_2 \leq N-1$ and $V \in \mathbb{S}^n$, along the state trajectory of (1), we have

$$x_{k_1+1}^{\mathrm{T}} V x_{k_2+1} = z_{k_1}^{\mathrm{T}} \Theta(V) z_{k_2}, \tag{96}$$

where $z_k = [x_k^{\mathrm{T}}, u_k^{\mathrm{T}}]^{\mathrm{T}}$ and

$$\Theta(V) = \begin{bmatrix} A^{\mathrm{T}} \\ B^{\mathrm{T}}\end{bmatrix} V \begin{bmatrix} A & B\end{bmatrix} = \begin{bmatrix} \Theta_{xx}(V) & \Theta_{ux}^{\mathrm{T}}(V) \\ \Theta_{ux}(V) & \Theta_{uu}(V)\end{bmatrix}. \tag{97}$$

It follows from (96) that

$$x_{[1,N]}^{\mathrm{T}} V x_{[1,N]} = z_{[0,N-1]}^{\mathrm{T}} \Theta(V) z_{[0,N-1]}. \tag{98}$$

Assumption 2 The exploration signal $u_{[0,N-1]}$ is PE of order $n+1$.

Under Assumption 2 and according to Lemma 12, $z_{[0,N-1]}$ is full row rank. As a result, for any fixed $V \in \mathbb{S}^n$, (98) admits a unique solution

$$\Theta(V) = \Lambda V \Lambda^{\mathrm{T}}, \tag{99}$$

where Λ is a data-dependent matrix defined as

$$(z_{[0,N-1]}^{\mathrm{T}})^{\dagger} = (z_{[0,N-1]} z_{[0,N-1]}^{\mathrm{T}})^{-1} z_{[0,N-1]}, \tag{100a}$$

$$\Lambda = (z_{[0,N-1]}^{\mathrm{T}})^{\dagger} x_{[1,N]}^{\mathrm{T}}. \tag{100b}$$

Therefore, given any $V \in \mathbb{S}^n$, $\Theta(V)$ can be directly computed from (99) without knowing the system matrices A and B.

By (97), we can rewrite (7) as

$$[I_n \; -L_i^{\mathrm{T}}] \, \Theta(V_i) \, [I_n \; -L_i^{\mathrm{T}}]^{\mathrm{T}} - V_i + Q + L_i^{\mathrm{T}} R L_i = 0. \tag{101}$$

Plugging (99) into (101) yields (102). The learning-based PI is represented in the following procedure.

Procedure 5 (Learning-based PI for discrete-time LQR)

1. *Learning-based policy evaluation*

$$\begin{aligned} &[I_n \; -L_i^{\mathrm{T}}] \, \Lambda V_i \Lambda^{\mathrm{T}} \, [I_n \; -L_i^{\mathrm{T}}]^{\mathrm{T}} - V_i + Q \\ &+ L_i^{\mathrm{T}} R L_i = 0. \end{aligned} \tag{102}$$

2. *Learning-based policy improvement*

$$L_{i+1} = (R + \Theta_{uu}(V_i))^{-1} \Theta_{ux}(V_i). \tag{103}$$

It should be noticed that due to (99), Procedure 5 is equivalent to Procedure 1.

4.2 Robustness analysis

In the previous subsection, we assume that the accurate data from system can be obtained. In reality, measurement noise and unknown system disturbance are inevitable. Therefore, the input-state data is sampled from the following linear system with unknown system disturbance and measurement noise

$$\begin{cases} \check{x}_{k+1} = A\check{x}_k + Bu_k + w_k, \\ y_k = \check{x}_k + v_k, \end{cases} \tag{104}$$

where $w_k \sim \mathcal{N}(0, \Sigma_w)$ and $v_k \sim \mathcal{N}(0, \Sigma_v)$ are independent and identically distributed random noises. Let $\check{z}_k = [y_k^{\mathrm{T}}, u_k^{\mathrm{T}}]$ and suppose there are in total S trajectories of system (104) which start from the same initial state and are driven by the same exploration input $u_{[0,N-1]}$. Averaging the collected data over S trajectories, we have

$$\bar{z}^S_{[0,N-1]} = \frac{1}{S} \sum_{s=1}^{S} \check{z}^s_{[0,N-1]}, \;\; \bar{y}^S_{[1,N]} = \tfrac{1}{S} \sum_{s=1}^{S} y^s_{[1,N]}. \tag{105}$$

Then, the data-dependent matrix is constructed as

$$\check{\Lambda}^S = [(\bar{z}^S_{[0,N-1]})^{\mathrm{T}}]^{\dagger} \bar{y}^S_{[1,N]}. \tag{106}$$

By the strong law of large numbers, the following limitations hold almost surely

$$\begin{aligned} &\lim_{S\to\infty} \bar{z}^S_{[0,N-1]} = z_{[0,N-1]}, \;\; \lim_{S\to\infty} \bar{y}^S_{[1,N]} = x_{[1,N]}, \\ &\lim_{S\to\infty} \check{\Lambda}^S = \Lambda. \end{aligned} \tag{107}$$

Recall that $z_{[0,N-1]}$, $x_{[1,N-1]}$, and Λ are the data collected from system (1) with the same initial state and exploration input as (104). Since S is finite, the difference between Λ and $\check{\Lambda}^S$ is unavoidable, and hence,

$$\check{\Lambda}^S = \Lambda + \Delta\Lambda^S. \tag{108}$$

Consequently, $\check{\Theta}(V) = \check{\Lambda}^S V (\check{\Lambda}^S)^{\mathrm{T}}$ is the estimation of $\Theta(V)$. Using the noisy data-dependent matrix $\check{\Lambda}^S$, the learning-based PI is presented as:

Procedure 6 (Learning-based PI using noisy data)

1. *Learning-based policy evaluation using noisy data*

$$\left[I_n \; -\check{L}_i^{\mathrm{T}}\right] \check{\Theta}(\check{V}_i) \left[I_n \; -\check{L}_i^{\mathrm{T}}\right]^{\mathrm{T}} - \check{V}_i + Q + \check{L}_i^{\mathrm{T}} R \check{L}_i = 0. \tag{109}$$

2. *Learning-based policy improvement using noisy data*

$$\check{L}_{i+1} = (R + \check{\Theta}_{uu}(\check{V}_i))^{-1} \check{\Theta}^{\mathrm{T}}_{xu}(\check{V}_i). \tag{110}$$

In Procedure 6, the symbol "check" is used to denote the variables for the learning-based PI using noisy data. In addition, let $\tilde{V}_i$ denote the result of the accurate evaluation of $\check{L}_i$, i.e. $\tilde{V}_i$ is the solution of (109) with $\check{\Theta}(\check{V}_i)$ replaced by $\Theta(\tilde{V}_i)$. $\check{V}_i = \tilde{V}_i + \Delta V_i$ is the solution of (109) and ΔV_i is the policy evaluation error induced by the noise $\Delta\Lambda$. In the following contents, the superscript S is omitted to simplify the notation. Based on the robustness analysis in the previous section, we will analyze the robustness of the learning-based PI to the noise $\Delta\Lambda$.

For any stabilizing control gain L, let $\check{V}_L = V_L + \Delta V$ be the solution of the learning-based policy evaluation with the noisy data-dependent matrix $\check{\Lambda}$, i.e.

$$\begin{aligned} &[I_n \; -L^{\mathrm{T}}] (\Lambda + \Delta\Lambda)(V_L + \Delta V)(\Lambda + \Delta\Lambda)^{\mathrm{T}} [I_n \; -L^{\mathrm{T}}]^{\mathrm{T}} \\ &- (V_L + \Delta V) + Q + L^{\mathrm{T}} R L = 0. \end{aligned} \tag{111}$$

The following lemma guarantees that (111) has a unique solution $(V_L + \Delta V) = (V_L + \Delta V)^{\mathrm{T}} \succ 0$.

Lemma 13 *If*

$$\|\Delta\Lambda\| < -\|\Lambda\| + \sqrt{\|\Lambda\|^2 + \frac{\underline{\lambda}(Q)}{(1 + \|L\|)^2 \|V_L\|}}, \tag{112}$$

then $(\Lambda + \Delta\Lambda)^{\mathrm{T}} \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}}$ *is a Schur matrix.*

Proof Recall that $V_L = V_L^{\mathrm{T}} \succ 0$ is the solution of (5) associated with the stabilizing control gain L. By (99), (5) is equivalent to the following equation

$$\left[I_n \ -L^{\mathrm{T}}\right] \Lambda V_L \Lambda^{\mathrm{T}} \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}} - V_L + Q + L^{\mathrm{T}} R L = 0. \tag{113}$$

Since $Q \succ 0$, by [30, Lemma 3.9], $\Lambda^{\mathrm{T}} \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}}$ is Schur.

When Λ is disturbed by $\Delta\Lambda$, we can rewrite (113) as

$$\begin{aligned}&\left[I_n \ -L^{\mathrm{T}}\right](\Lambda + \Delta\Lambda) V_L (\Lambda + \Delta\Lambda)^{\mathrm{T}} \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}} \\ &\quad - V_L - \left[I_n \ -L^{\mathrm{T}}\right](\Lambda V_L \Delta\Lambda^{\mathrm{T}} + \Delta\Lambda V_L \Lambda^{\mathrm{T}} \\ &\quad + \Delta\Lambda V_L \Delta\Lambda^{\mathrm{T}}) \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}} + Q + L^{\mathrm{T}} R L = 0.\end{aligned} \tag{114}$$

When (112) holds, we have

$$\begin{aligned}&\left[I_n \ -L^{\mathrm{T}}\right](\Lambda V_L \Delta\Lambda^{\mathrm{T}} + \Delta\Lambda V_L \Lambda^{\mathrm{T}} \\ &\quad + \Delta\Lambda V_L \Delta\Lambda^{\mathrm{T}}) \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}} - Q \prec 0.\end{aligned} \tag{115}$$

By [30, Lemma 3.9], (114) and (115), $(\Lambda + \Delta\Lambda)^{\mathrm{T}} \times \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}}$ is Schur. □

The following lemma implies that the policy evaluation error ΔV is small as long as $\Delta\Lambda$ is sufficiently small.

Lemma 14 *For any* $h > 0$, $L \in \mathcal{L}_h$, *and* $\Delta\Lambda$ *satisfying* (112), *we have*

$$\begin{aligned}&\frac{\|\Delta V\|}{\|V_L + \Delta V\|} \\ &\quad \le \frac{h + \mathrm{Tr}\,(V^*)}{\underline{\lambda}(Q)} (2\|\Lambda\| + \|\Delta\Lambda\|)(1 + \|L\|)^2 \|\Delta\Lambda\|.\end{aligned} \tag{116}$$

Proof According to [32, Theorems 2.6 and 4.1], we have

$$\begin{aligned}&\frac{\|\Delta V\|}{\|V_L + \Delta V\|} \\ &\quad \le \Big\| \sum_{k=0}^{\infty} (\left[I_n \ -L^{\mathrm{T}}\right] \Lambda)^k (\Lambda^{\mathrm{T}} \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}})^k \Big\| \\ &\quad\quad \times (2\|\Lambda\| + \|\Delta\Lambda\|)(1 + \|L\|)^2 \|\Delta\Lambda\|.\end{aligned} \tag{117}$$

Since $\mathrm{Tr}\,(V_L) \le h + \mathrm{Tr}\,(V^*)$, it follows from (102) and [27, Theorem 5.D6] that

$$V_L = \sum_{k=0}^{\infty} (\left[I_n \ -L^{\mathrm{T}}\right] \Lambda)^k (Q + L^{\mathrm{T}} R L)(\Lambda^{\mathrm{T}} \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}})^k. \tag{118}$$

Taking the trace of both sides of (118) and utilizing [28], we have

$$\begin{aligned}&h + \mathrm{Tr}\left(V^*\right) \\ &\quad \ge \underline{\lambda}(Q) \Big\| \sum_{k=0}^{\infty} (\left[I_n \ -L^{\mathrm{T}}\right] \Lambda)^k (\Lambda^{\mathrm{T}} \left[I_n \ -L^{\mathrm{T}}\right]^{\mathrm{T}})^k \Big\|.\end{aligned} \tag{119}$$

Plugging (119) into (117) yields (116). □

The following lemma tells us that $\Delta\Theta$ is small if ΔV and $\Delta\Lambda$ are small enough.

Lemma 15 *Let* $\check{\Theta}(\check{V}_L) = \check{\Lambda} \check{V}_L \check{\Lambda}^{\mathrm{T}}$ *and* $\Delta\Theta(V_L) = \check{\Theta}(\check{V}_L) - \Theta(V_L)$, *then,*

$$\begin{aligned}\|\Delta\Theta(V_L)\| &\le 2\|\Lambda\| \|V_L\| \|\Delta\Lambda\| + \|\Lambda\|^2 \|\Delta V\| \\ &\quad + 2\|\Lambda\| \|\Delta V\| \|\Delta\Lambda\| + \|V_L\| \|\Delta\Lambda\|^2 \\ &\quad + \|\Delta\Lambda\|^2 \|\Delta V\|.\end{aligned} \tag{120}$$

Proof By the expressions of $\check{\Theta}(\check{V}_L)$ and $\Theta(V_L)$, we have

$$\begin{aligned}\Delta\Theta(V_L) &= \Lambda V_L \Delta\Lambda^{\mathrm{T}} + \Lambda \Delta V_L \Lambda^{\mathrm{T}} + \Lambda \Delta V_L \Delta\Lambda^{\mathrm{T}} \\ &\quad + \Delta\Lambda V_L \Lambda^{\mathrm{T}} + \Delta\Lambda V_L \Delta\Lambda^{\mathrm{T}} + \Delta\Lambda \Delta V_L \Lambda^{\mathrm{T}} \\ &\quad + \Delta\Lambda \Delta V_L \Delta\Lambda^{\mathrm{T}}.\end{aligned} \tag{121}$$

Hence, (120) is obtained by (121) and the triangle inequality. □

By the follow lemma, it is ensured that ΔL converges to zero as $\Delta\Theta$ tends to zero.

Lemma 16 *Let* $\Delta L = (R + \check{\Theta}_{uu}(\check{V}_L))^{-1} \check{\Theta}_{ux}(\check{V}_L) - (R + \Theta_{uu}(V_L))^{-1} \Theta_{ux}(V_L)$. *Then,*

$$\|\Delta L\| \le \left(\underline{\lambda}(R)^{-1} + \underline{\lambda}(R)^{-2} \|\Theta(V_L)\|\right) \|\Delta\Theta(V_L)\|. \tag{122}$$

Proof From the expression of ΔL, we have

$$\begin{aligned}\Delta L &= (R + \check{\Theta}_{uu}(\check{V}_L))^{-1} \Delta\Theta_{ux}(V_L) + [(R + \check{\Theta}_{uu}(\check{V}_L))^{-1} \\ &\quad - (R + \Theta_{uu}(V_L))^{-1}] \Theta_{ux}(V_L) \\ &= (R + \check{\Theta}_{uu}(\check{V}_L))^{-1} \Delta\Theta_{ux}(V_L) \\ &\quad - (R + \check{\Theta}_{uu}(\check{V}_L))^{-1} \Delta\Theta_{uu}(V_L)(R \\ &\quad + \Theta_{uu}(V_L))^{-1} \Theta_{ux}(V_L).\end{aligned} \tag{123}$$

Therefore, (122) readily follows from (123). □

Given the aforementioned lemmas, we are ready to show the main result on the robustness of the learning-based PI algorithm in Procedure 6.

Theorem 3 *For any $\epsilon > 0$, $h > 0$, and $\hat{L}_0 \in \mathcal{L}_h$, there exist $i^* \in \mathbb{Z}_+$ and $f^* > 0$, such that if $\|\Delta\Lambda\| < f^*$, $\|\check{V}_i - V^*\| < \epsilon$ $\forall i \geq i^*$.*

Proof At each iteration of Procedure 6, if Λ is not disturbed by noise, i.e. $\Delta\Lambda = 0$, the policy improvement is (103). Due to the influence of $\Delta\Lambda$, the control gain is updated by (110), which can be rewritten as

$$\check{L}_{i+1} = (R + \Theta_{uu}(\tilde{V}_i))^{-1}\Theta_{xu}^{\mathrm{T}}(\tilde{V}_i) + \Delta L_{i+1}, \tag{124}$$

where ΔL_{i+1} is

$$\begin{aligned}\Delta L_{i+1} =&(R + \check{\Theta}_{uu}(\check{V}_i))^{-1}\check{\Theta}_{xu}^{\mathrm{T}}(\check{V}_i)\\ &- (R + \Theta_{uu}(\tilde{V}_i))^{-1}\Theta_{xu}^{\mathrm{T}}(\tilde{V}_i).\end{aligned} \tag{125}$$

By Lemmas 13, 14, 15, and 16, $\|\Delta L_{i+1}\|$ tends to zero as $\|\Delta\Lambda\|$ tends to zero. Therefore, there exists f_1, such that if $\|\Delta\Lambda\| < f_1$, $\|\Delta L_{i+1}\| < b(h)$, where $b(h)$ is defined in Lemma 5. Hence, $\check{L}_i \in \mathcal{L}_h$ for any $i \in \mathbb{Z}_+$. By Lemma 6, we have

$$\|\tilde{V}_i - V^*\|_F \leq \kappa(\|\tilde{V}_0 - V^*\|_F, i) + \rho(\|\Delta L\|_\infty). \tag{126}$$

As a result, there exist $i^* \in \mathbb{Z}_+$ and $f_2 > 0$, such that if $\|\Delta\Lambda\| < f_2$, $\|\tilde{V}_i - V^*\|_F < \epsilon/2$ $\forall i \geq i^*$. Furthermore, by Lemma 14, there exists $f_3 > 0$ such that $\|\check{V}_i - \tilde{V}_i\|_F < \epsilon/2$ $\forall i \geq i^*$ as long as $\|\Delta\Lambda\| < f_3$. In summary, by triangle inequality, if $\|\Delta\Lambda\| < f^* = \min\{f_1, f_2, f_3\}$, we have $\|\check{V}_i - V^*\|_F < \epsilon$ $\forall i \geq i^*$. □

Corollary 1 *For any $\epsilon > 0$, $h > 0$, and $\hat{L}_0 \in \mathcal{L}_h$, there exist $i^* \in \mathbb{Z}_+$ and $S^* \in \mathbb{Z}_+$, such that if $S > S^*$, $\|\check{V}_i - V^*\| < \epsilon$ $\forall i \geq i^*$.*

Proof For the given $f^* > 0$ in Theorem 3, by (107), there exists $S^* \in \mathbb{Z}_+$, such that if $S > S^*$, it is ensured that $\|\Delta\Lambda\| < f^*$. Then, the proof is completed by Theorem 3. □

5 Numerical simulation

In this section, we illustrate the proposed theoretical results by a benchmark example known as cart-pole system [33]. The parameters of the cart-pole system are: $m_c = 1$ kg (mass of the cart), $m_p = 0.1$ kg (mass of the pendulum), $l_p = 0.5$ m (distance from the center of mass of the pendulum to the pivot), $g_c = 9.8$ m/s^2 (gravitational acceleration). By linearizing the system around the equilibrium, the system is

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -0.717 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 15.77 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0.98 \\ 0 \\ -1.46 \end{bmatrix} u. \tag{127}$$

By discretizing it through Euler method with a step size of $0.01sec$, we have

$$x_{k+1} = \begin{bmatrix} 1 & 0.01 & 0 & 0 \\ 0 & 1 & -0.007 & 0 \\ 0 & 0 & 1 & 0.01 \\ 0 & 0 & 0.158 & 1 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0.0098 \\ 0 \\ -0.0146 \end{bmatrix} u_k. \tag{128}$$

The weighting matrices of the cost (1) are $Q = 10I_4$ and $R = 1$. The initial stabilizing gain to start the policy iteration algorithm is

$$\hat{K}_0 = \hat{L}_0 = \begin{bmatrix} -58.6 & -43.6 & -167.8 & -43.5 \end{bmatrix}. \tag{129}$$

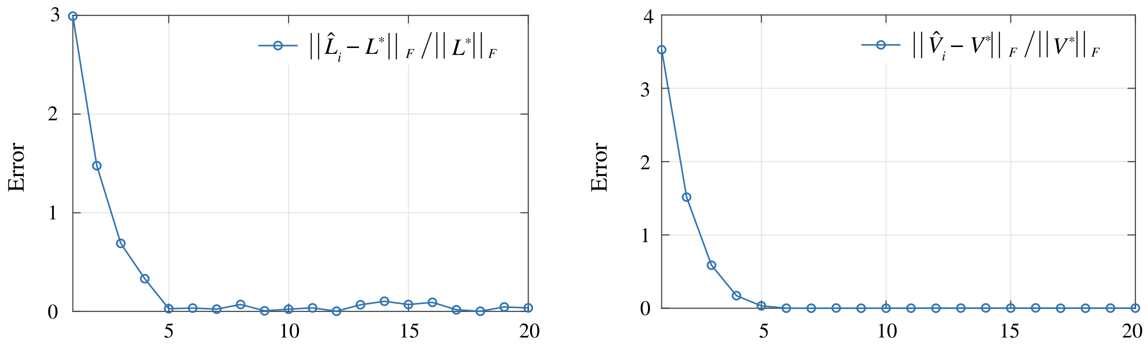
5.1 Robustness test of the inexact policy iteration

We test the robustness of the inexact PI for discrete-time systems in Procedure 3. At each iteration, each element of ΔG_i is sampled from a standard Gaussian distribution, and then its spectral norm is scaled to 0.2. During the iteration, the relative errors of the control gain $\hat{L}_i$ and cost matrix $\hat{V}_i$ are shown in Fig. 1. The control gain and cost matrix are close to the optimal solution at the 5th iteration. It is observed that even under the influence of disturbances at each iteration, the inexact PI in Procedure 3 can still approach the optimal solution. This is consistent with the ISS property of Procedure 3 in Theorem 1.

In addition, the robustness of Procedure 4 is tested. At each iteration, ΔM_i is randomly sampled with the norm of 0.2. Under the influence of ΔM_i, the evolution of the control gain $\hat{K}_i$ and cost matrix $\hat{P}_i$ is shown in Fig. 2. Under the noise ΔM_i, the algorithm cannot converge exactly to the optimal solution. However, with the small-disturbance ISS property, the inexact PI can still converge to a neighborhood of the optimal solution, which is consistent with Theorem 2.

5.2 Robustness test of the learning-based policy iteration

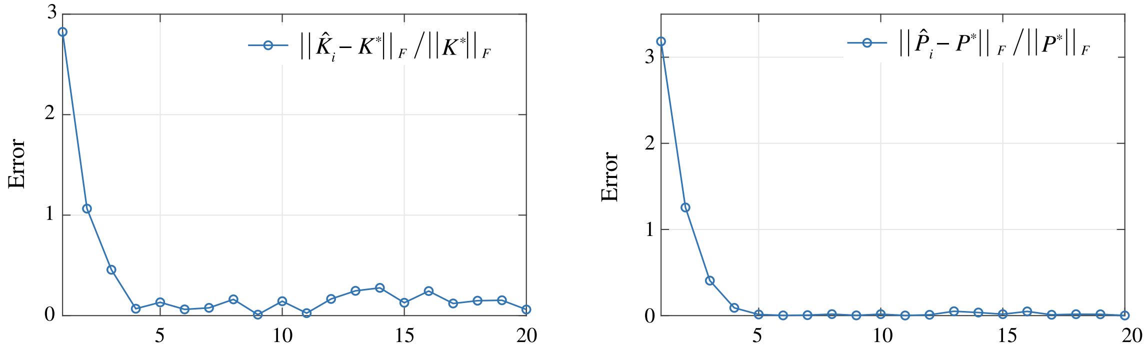
The robustness of the learning-based PI in Procedure 6 is tested for system (104) with both system disturbance and measurement noise. The variances of the system disturbance and measurement noise are $\Sigma_w = 0.01I_n$ and $\Sigma_v = 0.01I_n$. One trajectory is sampled from the solution of (104) and the length of the sampled trajectory is $N = 100$, i.e. 100 data collected from (104) is used to construct the data-dependent matrix $\hat{\Lambda}^S$. Compared with the matrix Λ in (100b) where the data is collected from the system without unknown system disturbance and measurement noise, $\hat{\Lambda}^S$ is directly constructed by the noisy data. Therefore, at each iteration of the learning-based PI, $\Delta\Lambda^S$ introduces the disturbances. The evolution of the control gain and cost matrix is in Fig. 3. It



(a) Control gain error

(b) Cost matrix error

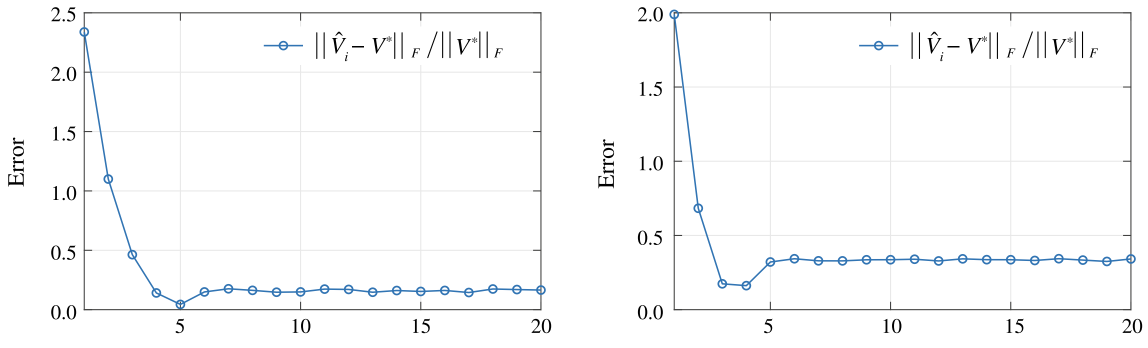
Fig. 1 Robustness test of Procedure 3 when $\|\Delta G\|_\infty = 0.2$



(a) Control gain error

(b) Cost matrix error

Fig. 2 Robustness test of Procedure 4 when $\|\Delta M\|_\infty = 0.2$



(a) Control gain error

(b) Cost matrix error

Fig. 3 Robustness test of Procedure 6 when the noisy data is applied for the construction of $\hat{\Lambda}$

is observed that with the noisy data, the control gain and the cost matrix obtained by Procedure 6 converge to an approximation of the optimal solution. This coincides with the result in Theorem 3.

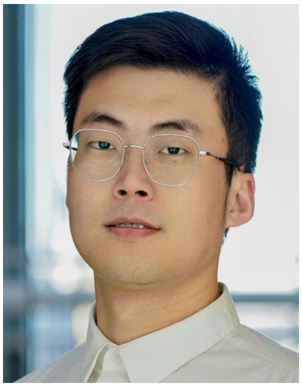
6 Conclusion

In this paper, we have studied the robustness property of policy optimization in the presence of disturbances at each iteration. Using ISS Lyapunov techniques, it is demonstrated that the PI ultimately converges to a small neighborhood of the optimal solution as long as the disturbance is sufficiently small. In this paper, we also provided a quantifiable bound. Based on the ISS property and Willems' fundamental lemma, a learning-based PI algorithm is proposed and the persist excitation of the exploratory signal can be easily guaranteed. A numerical simulation example is provided to illustrate the theoretical results.

Data availability The data that support the findings of this study are available from the corresponding author, L. Cui, upon reasonable request.

References

1. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
2. Fazel, M., Ge, R., Kakade, S., & Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th international conference on machine learning.* vol. 80, pp. 1467–1476.
3. Bu, J., Mesbahi, A., Fazel, M.,& Mesbahi, M. (2019). LQR through the lens of first order methods: Discrete-time case. arXiv:1907.08921 (arXiv e-preprint).
4. Hu, B., Zhang, K., Li, N., Mesbahi, M., Fazel, M., & Başar, T. (2022). Towards a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems, 6*(1), 123–158. https://doi.org/10.1146/annurev-control-042920-020021
5. Mohammadi, H., Zare, A., Soltanolkotabi, M., & Jovanovic, M. R. (2022). Convergence and sample complexity of gradient methods for the model-free linear-quadratic regulator problem. *IEEE Transactions on Automatic Control, 67*(5), 2435–2450.
6. Kleinman, D. (1968). On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control, 13*(1), 114–115. https://doi.org/10.1109/TAC.1968.1098829
7. Hewer, G. (1971). An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control, 16*(4), 382–384. https://doi.org/10.1109/TAC.1971.1099755
8. Bertsekas, D. P. (1995). *Dynamic programming and optimal control* (Vol. 2). Athena Scientific.
9. Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.
10. Jiang, Y., & Jiang, Z. P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica, 48*(10), 2699–2704. https://doi.org/10.1016/j.automatica.2012.06.096
11. Jiang, Y., & Jiang, Z. P. (2015). Global adaptive dynamic programming for continuous-time nonlinear systems. *IEEE Transactions on Automatic Control, 60*(11), 2917–2929. https://doi.org/10.1109/TAC.2015.2414811
12. Cui, L., Pang, B., & Jiang, Z. P. (2023). Learning-based adaptive optimal control of linear time-delay systems: A policy iteration approach. *IEEE Transactions on Automatic Control*. https://doi.org/10.1109/TAC.2023.3273786
13. Pang, B., Jiang, Z. P., & Mareels, I. (2020). Reinforcement learning for adaptive optimal control of continuous-time linear periodic systems. *Automatica, 118*, 109035. https://doi.org/10.1016/j.automatica.2020.109035
14. Gao, W., & Jiang, Z. P. (2016). Adaptive dynamic programming and adaptive optimal output regulation of linear systems. *IEEE Transactions on Automatic Control, 61*(12), 4164–4169. https://doi.org/10.1109/TAC.2016.2548662
15. Pang, B., Cui, L., & Jiang, Z. P. (2022). Human motor learning is robust to control-dependent noise. *Biological Cybernetics, 116*(12), 307–325.
16. Liu, T., Cui, L., Pang, B., & Jiang, Z. P. (2021). Data-driven adaptive optimal control of mixed-traffic connected vehicles in a ring road. In *2021 60th IEEE conference on decision and control (CDC)*, pp. 77–82. https://doi.org/10.1109/CDC45484.2021.9683024.
17. Cui, L., Ozbay, K., & Jiang, Z. P. (2021) Combined longitudinal and lateral control of autonomous vehicles based on reinforcement learning. In: *2021 American control conference (ACC)*, pp. 1929–1934. https://doi.org/10.23919/ACC50511.2021.9483388.
18. Ljung, L. (1998). *System identification* (pp. 163–173). Birkhäuser. https://doi.org/10.1007/978-1-4612-1768-8_11
19. Jiang, Z. P., Bian, T., & Gao, W. (2020). Learning-based control: A tutorial and some recent results. *Foundations and Trends in Systems and Control, 8*(3), 176–284. https://doi.org/10.1561/2600000023
20. Cui, L., Başar, T., & Jiang, Z. P. (2022). A reinforcement learning look risk-sensitive linear quadratic Gaussian control. In *5th Annual Learning for Dynamics and Control Conference*, pp. 534–546.
21. Pang, B., & Jiang, Z. P. (2023). Reinforcement learning for adaptive optimal stationary control of linear stochastic systems. *IEEE Transactions on Automatic Control, 68*(4), 2383–2390. https://doi.org/10.1109/TAC.2022.3172250
22. Cui, L., Wang, S., Zhang, J., Zhang, D., Lai, J., Zheng, Y., Zhang, Z., & Jiang, Z. P. (2021). Learning-based balance control of wheel-legged robots. *IEEE Robotics and Automation Letters, 6*(4), 7667–7674. https://doi.org/10.1109/LRA.2021.3100269
23. Sontag, E. (2008). *Input to state stability: Basic concepts and results. Lecture notes in mathematics* (pp. 163–220). Springer.
24. Pang, B., & Jiang, Z. P. (2021). Robust reinforcement learning: A case study in linear quadratic regulation. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(10), 9303–9311.
25. Pang, B., Bian, T., & Jiang, Z. P. (2022). Robust policy iteration for continuous-time linear quadratic regulation. *IEEE Transactions on Automatic Control, 67*(1), 504–511. https://doi.org/10.1109/TAC.2021.3085510
26. Chen, B. M. (2013). In J. Baillieul & T. Samad (Eds.), *H2 optimal control.* Springer. https://doi.org/10.1007/978-1-4471-5102-9_204-1
27. Chen, C.-T. (1999). *Linear system theory and design*. Oxford University Press.
28. Mori, T. (1988). Comments on "A matrix inequality associated with bounds on solutions of algebraic Riccati and Lyapunov equation" by J. M. Saniuk and I.B. Rhodes. *IEEE Transactions on Automatic Control, 33*(11), 1088. https://doi.org/10.1109/9.14428
29. Hespanha, J. P. (2018). *Linear systems theory*. Princeton University Press.
30. Zhou, K., Doyle, J. C., & Glover, K. (1996). *Robust and optimal control*. Prentice Hall.
31. Willems, J. C., Rapisarda, P., Markovsky, I., & De Moor, B. L. M. (2005). A note on persistency of excitation. *Systems and Control Letters, 54*(4), 325–329. https://doi.org/10.1016/j.sysconle.2004.09.003
32. Gahinet, P. M., Laub, A. J., Kenney, C. S., & Hewer, G. A. (1990). Sensitivity of the stable discrete-time Lyapunov equation. *IEEE Transactions on Automatic Control, 35*(11), 1209–1217. https://doi.org/10.1109/9.59806
33. Anderson, C. W. (1989). Learning to control an inverted pendulum using neural networks. *IEEE Control Systems Magazine, 9*(3), 31–37. https://doi.org/10.1109/37.24809



Leilei Cui received the B.Eng. degree in automation from Northwestern Polytechnical University, Xian, China, in 2016, and the M.Sc. degree in control engineering from Shanghai Jiao Tong University, Shanghai, China, in 2019. He is currently a Ph.D. candidate in the Control and Networks Lab, Tandon School of Engineering, New York University. His research interests include robot control, reinforcement learning, adaptive dynamic programming, and optimal control.

Zhong-Ping Jiang received the M.Sc. degree in statistics from the University of Paris XI, France, in 1989, and the Ph.D. degree in automatic control and mathematics from the Ecole des Mines de Paris (now, called ParisTech-Mines), France, in 1993, under the direction of Prof. Laurent Praly. Currently, he is a Professor of Electrical and Computer Engineering at the Tandon School of Engineering, New York University. His main research interests include stability theory, robust/adaptive/distributed nonlinear control, robust adaptive dynamic programming, reinforcement learning and their applications to information, mechanical and biological systems. In these fields, he has written six books and is author/co-author of over 500 peer-reviewed journal and conference papers. Prof. Jiang is a recipient of the prestigious Queen Elizabeth II Fellowship Award from the Australian Research Council, CAREER Award from the U.S. National Science Foundation, JSPS Invitation Fellowship from the Japan Society for the Promotion of Science, Distinguished Overseas Chinese Scholar Award from the NSF of China, and several best paper awards. He has served as Deputy Editor-in-Chief, Senior Editor and Associate Editor for numerous journals. Prof. Jiang is a Fellow of the IEEE, IFAC, CAA, and AAIA, a foreign member of the Academia Europaea (Academy of Europe), and is among the Clarivate Analytics Highly Cited Researchers. In 2022, he received the Excellence in Research Award from the NYU Tandon School of Engineering.