Robust Reinforcement Learning: A Case Study in Linear Quadratic Regulation
Bo Pang and Zhong-Ping Jiang

Abstract—This paper studies the robustness aspect of reinforcement learning algorithms in the presence of errors. Specifically, we revisit the benchmark problem of discrete-time linear quadratic regulation (LQR) and study the long-standing open question: Under what conditions is the policy iteration method robustly stable for dynamical systems with unbounded, continuous state and action spaces? Using advanced stability results in control theory, it is shown that policy iteration for LQR is inherently robust to small errors and enjoys local input-to-state stability: whenever the error in each iteration is bounded and small, the solutions of the policy iteration algorithm are also bounded, and, moreover, enter and stay in a small neighborhood of the optimal LQR solution. As an application, a novel off-policy optimistic least-squares policy iteration for the LQR problem is proposed, when the system dynamics are subjected to additive stochastic disturbances. The proposed new results in robust reinforcement learning are validated by a numerical example.

I. INTRODUCTION

As an important and popular method in reinforcement learning (RL), policy iteration has been widely studied by researchers and utilized in different kinds of real-life applications by practitioners

[6], [40]. Policy iteration involves two steps, policy evaluation and policy improvement. In policy evaluation, a given policy is evaluated based on a scalar performance index. Then this performance index is utilized to generate a new control policy in policy improvement. These two steps are iterated in turn, to find the solution of the RL problem. At hand. When all the information involved in this process is exactly known, the convergence to the optimal solution can be provably guaranteed, by exploiting the monotonicity property of the policy improvement step. That is, the performance of the newly generated policy is no worse than that of the given policy in each iteration. Over the past decades, various versions of policy iteration have been proposed, for diverse optimal control problems, see [6], [40], [24], [18] and the references therein.

In reality, policy evaluation or policy improvement can hardly be implemented precisely, because of the existence of various errors, which may be induced by function approximation, state estimation, sensor noise, external disturbance and so on. Therefore, a natural question to ask is: when is a policy iteration algorithm robust in the presence of errors? In other words, under what conditions on the errors, does the policy iteration still converge (to a neighborhood of) the optimal solution? And how to quantify the size of this neighborhood? In spite of the popularity and empirical successes of policy iteration, its robustness issue has not been fully understood yet in theory, due to the inherent nonlinearity of the process [27]. The problem becomes more complex when the state and action spaces are unbounded and continuous, which are common in RL problems of physical systems such as robotics and autonomous cars [25]. Indeed, in this case the stability issue needs to be addressed, to avoid the selection of destabilizing policies that drive the states of the closed-loop system into the infinity.

In this paper, we investigate the robustness of policy iteration for the discrete-time linear quadratic regulator (LQR) problem, which was firstly proposed in [15]. Even if the LQR is the most basic and important optimal control problem with unbounded, continuous state and action spaces [4], the robustness of its associated policy iteration has not been fully investigated. The main idea of this paper is to regard the policy iteration as a dynamical system, and then utilize the concepts of exponential stability and input-to-state stability in control theory to analyze its robustness [39]. To be more specific, we firstly prove that the optimal LQR solution is a locally exponentially stable equilibrium of the exact policy iteration. Then based on this observation, we show that the policy iteration with errors is locally input-to-state stable, if the errors are regarded as the control input. That is, if the policy iteration starts from an initial solution close to the optimal solution, and the errors are small and bounded, the discrepancies between the solutions generated by the policy iteration and the optimal solution will also be small and bounded. Thirdly, we demonstrate that for any initial stabilizing control gain, as long as the errors are small, the approximate solution given by policy iteration will eventually enter a small neighbourhood of the optimal solution. Finally, a novel off-policy model-free RL algorithm, named optimistic least-squares policy iteration (O-LSPI), is proposed for the LQR problem with dynamics perturbed by additive stochastic disturbances. Our robustness result is applied to show the convergence of this off-policy O-LSPI. To the best knowledge of the authors, no previous convergence results of off-policy RL algorithms are reported for such a LQR problem. Experiments on a numerical example validate our results.

Our main contributions are two-fold. First, we provide a control-theoretic robustness analysis for the policy iteration of discrete-time LQR. Second, we propose a novel off-policy RL algorithm O-LSPI with provable convergence.

In the rest of this paper, we first present some preliminaries, followed by the robustness analysis and the off-policy O-LSPI. Then we present the experimental results, discuss some related work, and close the paper with some concluding remarks.

A. Notations.

\( \mathbb{R} \) (\( \mathbb{R}_+ \)) is the set of all real (nonnegative) numbers; \( \mathbb{Z}_+ \) denotes the set of nonnegative integers; \( \mathbb{S}^n \) is the set of all real symmetric matrices of order \( n \); \( \otimes \) denotes the Kronecker product; \( I_n \) denotes the identity matrix with dimension \( n \); \( \| \cdot \|_F \) is the Frobenius norm; \( \| \cdot \|_2 \) is the 2-norm for vectors and the induced 2-norm for matrices; for signal \( Z : F \to \mathbb{R}^{m \times n} \), \( \| Z \|_\infty \) denotes its \( \ell^\infty \)-norm when \( F = \mathbb{Z}_+ \), and \( \ell^\infty \)-norm when \( F = \mathbb{R}_+ \). For matrices \( X \in \mathbb{R}^{m \times n}, Y \in \mathbb{S}^n \), and vector \( v \in \mathbb{R}^n \), define

\[
\text{vec}(X) = [ X_1^T \quad X_2^T \quad \cdots \quad X_n^T ]^T,
\]

\[
\text{vec}(Y) = [ y_{11}, \sqrt{2} y_{12}, \cdots, \sqrt{2} y_{1m}, y_{22}, \sqrt{2} y_{23}, \cdots, \sqrt{2} y_{m-1,m}, y_{mm} ]^T \in \mathbb{R}^{\frac{1}{2}n(m+1)},
\]

\( \tilde{v} = \text{vec}(v^T) \),

\( svec(\cdot) \) is the set of all real (nonnegative) numbers; \( \mathbb{Z}_+ \) denotes the Kronecker product; \( I_n \) denotes the identity matrix with dimension \( n \); \( \| \cdot \|_F \) is the Frobenius norm; \( \| \cdot \|_2 \) is the 2-norm for vectors and the induced 2-norm for matrices; for signal \( Z : F \to \mathbb{R}^{m \times n} \), \( \| Z \|_\infty \) denotes its \( \ell^\infty \)-norm when \( F = \mathbb{Z}_+ \), and \( \ell^\infty \)-norm when \( F = \mathbb{R}_+ \). For matrices \( X \in \mathbb{R}^{m \times n}, Y \in \mathbb{S}^n \), and vector \( v \in \mathbb{R}^n \), define

\[
\text{vec}(X) = [ X_1^T \quad X_2^T \quad \cdots \quad X_n^T ]^T,
\]

\[
\text{vec}(Y) = [ y_{11}, \sqrt{2} y_{12}, \cdots, \sqrt{2} y_{1m}, y_{22}, \sqrt{2} y_{23}, \cdots, \sqrt{2} y_{m-1,m}, y_{mm} ]^T \in \mathbb{R}^{\frac{1}{2}n(m+1)},
\]

\( \tilde{v} = \text{vec}(v^T) \),

This work has been supported in part by the U.S. National Science Foundation under Grants ECCS-1501044 and EPCN-1903781.

B. Pang and Z. P. Jiang are with the Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, 370 Jay Street, Brooklyn, NY 11201, USA (e-mail: bo.pang@nyu.edu; zjiang@nyu.edu).
where $X_i$ is the $i$th column of $X$. For $Z \in \mathbb{R}^{m \times n}$, define $B_r(Z) = \{X \in \mathbb{R}^{m \times n} \mid X - Z \leq r \}$ and $B_r(Z)$ as the closure of $B_r(Z)$. $Z^\top$ is the Moore-Penrose inverse of matrix $Z$.

II. Preliminaries

Consider linear time-invariant systems of the form

$$x_{k+1} = Ax_k + Bu_k, \quad x_0 = x_{ini}$$

(1)

where $x \in \mathbb{R}^n$ is the system state, $u \in \mathbb{R}^m$ is the control input, $x_0 = x_{ini} \in \mathbb{R}^n$ is the initial condition, $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$. $(A, B)$ is controllable, that is, $[B, AB, A^2B, \ldots, A^{n-1}B]$ has full row rank. The classic LQR problem is to find a controller $u$ in order to minimize the following cost functional

$$J = \sum_{k=0}^{\infty} c(x_k, u_k),$$

(2)

where $c(x, u) = x^T S x + u^T R u$, $S \in \mathbb{S}^n$ is positive semidefinite and $R \in \mathbb{S}^n$ is positive definite. $(A, S^{1/2})$ is observable, that is, $(A^T, S^{1/2})$ is controllable. It is well-known that under such a setting, the LQR problem admits a unique optimal controller $u^* = -K^*x$, where

$$K^* = (R + B^T P^* B)^{-1}B^T P^* A$$

(3)

with $P^* \in \mathbb{S}^n$ the unique positive definite solution of the algebraic Riccati equation (ARE)

$$A^T P A - P - A^T P B (R + B^T P B)^{-1} B^T P A + S = 0.$$ (4)

In addition, $A - BK^*$ is stable, i.e., the spectral radius $\rho(A - BK^*) < 1$. See [24] Section 2.4 for details. For convenience, a control gain $K \in \mathbb{R}^{n \times m}$ is said to be stabilizing if $A - BK$ is stable.

A. Policy Iteration for LQR

For any stabilizing control gain $K \in \mathbb{R}^{n \times m}$, the cost (2) with $u_k = -Kx_k$ is a quadratic function of the initial state [24] Section 2.4. Specifically, $J(x_0, -Kx) = x_0^T P x_0$, where $P \in \mathbb{S}^n$ is the unique positive definite solution of the Lyapunov equation

$$(A - BK)^T P K (A - BK) - P K + S + K^T R K = 0.$$ (5)

Define function

$$G(P) = \begin{bmatrix} [G(P)]_{xx} & [G(P)]_{ux} \\ [G(P)]_{ux}^T & [G(P)]_{uu} \end{bmatrix} \triangleq \begin{bmatrix} S + A^T P K A - P K & A^T P K B \\ B^T P K A & R + B^T P K B \end{bmatrix}.$$ (6)

Then (5) can be rewritten as

$$\mathcal{H}(G(P), K) = 0,$$

where

$$\mathcal{H}(G(P), K) \triangleq \begin{bmatrix} I_n & -K^T \end{bmatrix} G(P) \begin{bmatrix} I_n \\ -K \end{bmatrix}.$$ (7)

The policy iteration for LQR is presented below, which is an equivalent reformulation of the original results in [15].

Procedure 1 (Exact Policy Iteration).

1) Choose a stabilizing control gain $K_i$, and let $i = 1$.

2) (Policy evaluation) Evaluate the performance of control gain $K_i$, by solving

$$\mathcal{H}(G_i, K_i) = 0$$

for $P_i \in \mathbb{S}^n$, where $G_i = G(P_i)$.

C. Notions of exponential and input-to-state stability

Consider a dynamical system of the general form

$$x_{k+1} = f(x_k, u_k), \quad x_0 = x_{ini},$$

(9)

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ is continuous, and $x^*$ is an equilibrium of $x_{k+1} = f(x_k, 0)$ when $u_k = 0$ for all $k \in \mathbb{Z}_+$. The concepts of exponential and input-to-state stability for (9) are recalled in this subsection. See [19] for more details.

Definition 1. For (9) with $u_k = 0$ for all $k \in \mathbb{Z}_+$, $x^*$ is a locally exponentially stable equilibrium if there exists a $\delta > 0$, such that for some $a > 0$ and $0 < b < 1$,

$$\|x_k - x^*\|_2 \leq ab^k \|x_{ini} - x^*\|_2$$

(10)
for all \( x_{ini} \in \mathcal{B}_L(x^*) \). If \( \delta = +\infty \), then \( x^* \) is a globally exponentially stable equilibrium.

The exponential stability implies not only the convergence, but also the convergence rate of (9). When the input signal is not zero, the input-to-state stability characterizes how the solution of (9) is affected by the input signal.

**Definition 2.** A function \( \gamma : \mathbb{R}_+ \to \mathbb{R}_+ \) is said to be of class \( \mathcal{K} \) if it is continuous, strictly increasing and vanishes at the origin. A function \( \beta : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+ \) is said to be of class \( \mathcal{KL} \) if \( \beta(t, \ell) \) is of class \( \mathcal{K} \) for every fixed \( t \in \mathbb{R}_+ \) and, for every fixed \( r \geq 0 \), \( \beta(r, \ell) \) decreases to 0 as \( t \to \infty \).

**Definition 3.** System (9) is locally input-to-state stable if there exist some \( \alpha_1 > 0 \), some \( \alpha_2 > 0 \), some \( \beta \in \mathcal{KL} \) and some \( \gamma \in \mathcal{K} \), such that for each \( u \) and each \( x_{ini} \) satisfying \( x_{ini} \in \mathcal{B}_{\alpha_1}(x^*) \), \( \|u\| < \alpha_2 \), the corresponding solution \( x_k \) satisfies

\[
\|x_k - x^*\|_2 \leq \beta(\|x_{ini} - x^*\|_2, k) + \gamma(\|u\|_\infty).
\]

**Theorem 2.** For and its associated \( \delta_0 \) in Lemma 1 there exists \( \delta_1(\delta_0) > 0 \), such that if \( \|\Delta G\|_\infty < \delta_1 \), \( P_0 \in \mathcal{B}_{\delta_0}(P^*) \),

\( i ) \) \( \mathcal{G}_{1, uu} \) is invertible, \( \forall i \in \mathbb{Z}_+, i > 0; \)

\( ii ) \) is locally input-to-state stable (see Definition 3):

\[
\|\hat{P}_i - P^*\|_F \leq \beta(\|\hat{P}_0 - P^*\|_F, i) + \gamma(\|\Delta G\|_\infty), \quad \forall i \in \mathbb{Z}_+,
\]

where

\[
\beta(y, i) = \sigma y, \quad \gamma(y) = \frac{c_3}{1 - \sigma}y, \quad y \in \mathbb{R},
\]

and \( c_3(\delta_0) > 0 \).

(iii) \( \|\hat{K}_i\|_F < \kappa_1 \) for some \( \kappa_1(\delta_0) \in \mathbb{R}_+ \), \( \forall i \in \mathbb{Z}_+, i > 0; \)

(iv) \( \lim_{i \to \infty} \|\Delta G\|_F = 0 \) implies \( \lim_{i \to \infty} \|\hat{P}_i - P^*\|_F = 0 \).

To prove Theorem 2, we firstly prove that with the given conditions, by continuity \( \mathcal{G}_{1, uu} \) is invertible, \( \hat{K}_i \) is stabilizing and

\[
\|E(\hat{G}_i, \Delta G_i)\|_F \leq c_3(\|\Delta G\|_\infty).\]

Then obviously

\[
\vec{L}_X(Y) = \mathcal{A}(X) \vec{Y}(Y).
\]

**Theorem 3.** System (9) is locally input-to-state stable if there exist some \( \alpha_1 > 0 \), some \( \alpha_2 > 0 \), some \( \beta \in \mathcal{KL} \) and some \( \gamma \in \mathcal{K} \), such that for each \( u \) and each \( x_{ini} \) satisfying \( x_{ini} \in \mathcal{B}_{\alpha_1}(x^*) \), \( \|u\| < \alpha_2 \), the corresponding solution \( x_k \) satisfies

\[
\|x_k - x^*\|_2 \leq \beta(\|x_{ini} - x^*\|_2, k) + \gamma(\|u\|_\infty).
\]

**Theorem 4.** For \( \sigma \) and its associated \( \delta_0 \) in Lemma 1 there exists \( \delta_1(\delta_0) > 0 \), such that if \( \|\Delta G\|_\infty < \delta_1 \), \( P_0 \in \mathcal{B}_{\delta_0}(P^*) \),

\( i ) \) \( \mathcal{G}_{1, uu} \) is invertible, \( \forall i \in \mathbb{Z}_+, i > 0; \)

\( ii ) \) is locally input-to-state stable (see Definition 3):

\[
\|\hat{P}_i - P^*\|_F \leq \beta(\|\hat{P}_0 - P^*\|_F, i) + \gamma(\|\Delta G\|_\infty), \quad \forall i \in \mathbb{Z}_+,
\]

where

\[
\beta(y, i) = \sigma y, \quad \gamma(y) = \frac{c_3}{1 - \sigma}y, \quad y \in \mathbb{R},
\]

and \( c_3(\delta_0) > 0 \).

(iii) \( \|\hat{K}_i\|_F < \kappa_1 \) for some \( \kappa_1(\delta_0) \in \mathbb{R}_+ \), \( \forall i \in \mathbb{Z}_+, i > 0; \)

(iv) \( \lim_{i \to \infty} \|\Delta G\|_F = 0 \) implies \( \lim_{i \to \infty} \|\hat{P}_i - P^*\|_F = 0 \).

To prove Theorem 2, we firstly prove that with the given conditions, by continuity \( \mathcal{G}_{1, uu} \) is invertible, \( \hat{K}_i \) is stabilizing and

\[
\|E(\hat{G}_i, \Delta G_i)\|_F \leq c_3(\|\Delta G\|_\infty).\]

Then obviously

\[
\vec{L}_X(Y) = \mathcal{A}(X) \vec{Y}(Y).
\]

**Theorem 5.** System (9) is locally input-to-state stable if there exist some \( \alpha_1 > 0 \), some \( \alpha_2 > 0 \), some \( \beta \in \mathcal{KL} \) and some \( \gamma \in \mathcal{K} \), such that for each \( u \) and each \( x_{ini} \) satisfying \( x_{ini} \in \mathcal{B}_{\alpha_1}(x^*) \), \( \|u\| < \alpha_2 \), the corresponding solution \( x_k \) satisfies

\[
\|x_k - x^*\|_2 \leq \beta(\|x_{ini} - x^*\|_2, k) + \gamma(\|u\|_\infty).
\]

**Theorem 6.** For \( \sigma \) and its associated \( \delta_0 \) in Lemma 1 there exists \( \delta_1(\delta_0) > 0 \), such that if \( \|\Delta G\|_\infty < \delta_1 \), \( P_0 \in \mathcal{B}_{\delta_0}(P^*) \),

\( i ) \) \( \mathcal{G}_{1, uu} \) is invertible, \( \forall i \in \mathbb{Z}_+, i > 0; \)

\( ii ) \) is locally input-to-state stable (see Definition 3):

\[
\|\hat{P}_i - P^*\|_F \leq \beta(\|\hat{P}_0 - P^*\|_F, i) + \gamma(\|\Delta G\|_\infty), \quad \forall i \in \mathbb{Z}_+,
\]

where

\[
\beta(y, i) = \sigma y, \quad \gamma(y) = \frac{c_3}{1 - \sigma}y, \quad y \in \mathbb{R},
\]

and \( c_3(\delta_0) > 0 \).

(iii) \( \|\hat{K}_i\|_F < \kappa_1 \) for some \( \kappa_1(\delta_0) \in \mathbb{R}_+ \), \( \forall i \in \mathbb{Z}_+, i > 0; \)

(iv) \( \lim_{i \to \infty} \|\Delta G\|_F = 0 \) implies \( \lim_{i \to \infty} \|\hat{P}_i - P^*\|_F = 0 \).

To prove Theorem 2, we firstly prove that with the given conditions, by continuity \( \mathcal{G}_{1, uu} \) is invertible, \( \hat{K}_i \) is stabilizing and

\[
\|E(\hat{G}_i, \Delta G_i)\|_F \leq c_3(\|\Delta G\|_\infty).\]

Then obviously

\[
\vec{L}_X(Y) = \mathcal{A}(X) \vec{Y}(Y).
\]
stability of Lemma 1 prevents the accumulated effects of disturbance $\Delta G_i$ from driving $\|P_{i+1} - P^*\|_F$ to the infinity.

Intuitively, Theorem 2 implies that if $P_0$ is near $P^*$ (thus $\hat{K}_1$ is near $K^*$), and the disturbance input $\Delta G$ is bounded and not too large, then the cost of the generated control policy $\hat{K}_i$ is also bounded, and will ultimately be no larger than a constant proportional to the $l^\infty$-norm of the disturbance. The smaller the disturbance is, the better the ultimately generated policy is. In other words, the algorithm described in Procedure 2 is not sensitive to small disturbances when the initial condition is in a neighbourhood of the optimal solution.

The requirement that the initial condition $P_0$ need be in a neighbourhood of $P^*$ in Theorem 2 can be removed, as stated in the following corollary whose proof is given in the Appendix D.

**Corollary 1.** For any given stabilizing control gain $K_1$, and any $\epsilon > 0$, if $S > 0$, there exit $\delta_2(s, K_1) > 0$, $\Pi(\delta_2) > 0$, and $\kappa(\delta_2) > 0$, such that as long as $\|\Delta G_i\|_\infty < \delta_2$, $G_i$ is invertible, $K_i$ is stabilizing, $\|\hat{P}_i\|_F < \Pi$, $\|\hat{K}_i\|_F < \kappa$, $\forall i \in \mathbb{Z}_+$, $\epsilon > 0$ and

$$\limsup_{i \to \infty} \|\hat{P}_i - P^*\|_F < \epsilon.$$ 

If in addition $\lim_{i \to \infty} \|\Delta G_i\|_F = 0$, then $\lim_{i \to \infty} \|\hat{P}_i - P^*\|_F = 0$.

Here are the essential elements of the proof for Corollary 1. It is firstly proved that given any stabilizing control gain $K_1$, there exist $\bar{i} \in \mathbb{Z}_+$, $\bar{i} < +\infty$, and $\bar{b}_i > 0$, such that if $\|\Delta G_i\|_F < \bar{b}_i$ for $i = 1, 2, \ldots, \bar{i}$, then the $G_i$ is invertible, $K_i$ is stabilizing and bounded, $\hat{P}_i$ is bounded, $i = 1, 2, \ldots, \bar{i}$, (2) $\hat{P}_i$ enters the neighbourhood of $P^*$, i.e., $B_{\bar{b}_i}(P^*)$ defined in Theorem 2. Secondly, an application of Theorem 2 completes the proof.

In Corollary 1, $K_1$ can be any stabilizing control gain, which is different from that of Theorem 2. When there is no stabilizing Corollary 1 implies the convergence result of Procedure 1 in [13, Theorem 1] (i.e. Theorem 1 in this paper).

IV. Optimistic Least-Squares Policy Iteration

For system (3), due to the presence of stochastic noise $w_k$, the cost function (2) will not be finite. Thus alternatively the objective is to find a control law in the form of $u = -Kx$ directly from the input/state data, minimizing the cost function

$$J_{Avg}(u) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} c(x_k, u_k),$$

where $S$ and $R$ in $c(x_k, u_k)$ are positive definite. It is well-known [6, Section 4.4] that this problem shares the same optimal solutions with the standard LQR for system (1) and cost function (5). Specifically, the optimal control gain is given by (5), and the optimal control is $u_{Avg} = \text{tr}(C^T P^* C)$, with $P^*$ the unique positive definite solution of (6). For any stabilizing gain $K$, the cost it induces is $J_{Avg}(-Kx) = \text{tr}(C^T P K C)$, with $P$ and $K$ satisfying (5) (or equivalently (6)). Note that the assumption that $w_k \sim N(0, I_q)$ in (5) is not a restriction, since any random variable $X \sim N(0, \Sigma)$ with $\Sigma \in \mathbb{S}^q$ positive semidefinite, can be represented by $X_1 = D X_2$, where $\Sigma = D^T D$, $D \in \mathbb{R}^{q \times q}$, and $X_2 \sim N(0, I_q)$. Then $D$ is absorbed into $C$ in (5).

The optimistic least-squares policy iteration (O-LSPI) is based on the following observation: for a stabilizing gain $K$, its associated $P_K$ is the stable equilibrium of linear dynamical system

$$P_{K,j+1} = \mathcal{H}(Q(P_{K,j}), K), \quad P_{K,0} \in \mathbb{R}^{n \times n},$$

where

$$Q(P_{K,j}) = \begin{bmatrix} [Q(P_{K,j})]_{xx} & [Q(P_{K,j})]_{ux}^T \\ [Q(P_{K,j})]_{ux} & [Q(P_{K,j})]_{uu} \end{bmatrix} \begin{bmatrix} S + A^T P_{K,j} A & A^T P_{K,j} B \\ B^T P_{K,j} A & R + B^T P_{K,j} B \end{bmatrix},$$

(19)

This fact can be easily verified by rewriting and vectorizing (18) into its equivalent form

$$p_{K,j+1} = \left( (A - BK)^T \otimes (A - BK)^T \right) p_{K,j} + \text{vec}(S + K^T R K), \quad p_{K,0} \in \mathbb{R}^{q^2},$$

(20)

where $p_{K,j} = \vec{p}(P_{K,j})$. Since $(A - BK)^T \otimes (A - BK)^T$ is also stable. Thus (20) admits a unique stable equilibrium. So does (18) and the unique solution must be $P_K$ because

$$Q(P_{K,j}) = G(P_{K,j}) + \begin{bmatrix} P_{K,j} & 0 \\ 0 & 0 \end{bmatrix},$$

(21)

$$\mathcal{H}(G(P_{K,j}), K) = \mathcal{H}(Q(P_{K,j}), K) - p_{K,j}.$$ This implies that instead of solving (6), we may utilize iteration (18) to achieve policy evaluation. It is not hard to recognize that (18) is actually the LQR version of the optimistic policy iteration in (22) for problems with discrete state and action spaces (thus the name “optimistic” in O-LSPI). Suppose a behavior policy $u_k = -K_1 x_k + v_k$ is applied to the system to collect data, where $K_1$ is stabilizing and $v_k$ is drawn i.i.d. from Gaussian distribution $N(0, \sigma^2_0 I_m)$, $\sigma_0 \in \mathbb{R}_+$. Then the state-control pair $[x^T, u^T]^T$ admits a unique invariant distribution $\pi$. We make the following assumption.

**Assumption 1.** $E_{\pi} \left[ z \tilde{z}^T \right]$ is invertible, where $z = [x^T, u^T, 1]^T$.

For any $P \in \mathbb{S}^n$, we have

$$E \left[ z_k z_k^T P x_{k+1} x_k, u_k \right] = E \left[ z_k^T F(P) z_k | x_k, u_k \right] - c(x_k, u_k)$$

where

$$F(P) = \begin{bmatrix} Q(P) & 0 \\ 0 & \text{tr}(C^T P C) \end{bmatrix}.$$ Vectorizing and multiplying the above equation by $\tilde{z}_k$ yields

$$E \left[ \tilde{z}_k \tilde{z}_k^T P x_{k+1} x_k, u_k \right] \text{svec}(P) = E \left[ \tilde{z}_k^T F(P) \tilde{z}_k | x_k, u_k \right] - \tilde{z}_k c(x_k, u_k).$$

Taking expectation with respect to the invariant distribution $\pi$, by Assumption 1 we obtain

$$\text{svec}(F(P)) = \varphi_1^{-1} \left( \varphi_2 \text{svec}(P) + \varphi_3 \right).$$ (22)

where

$$\varphi_1 = E_{\pi} \left[ \tilde{z}_k \tilde{z}_k^T \right],$$

$$\varphi_2 = E_{\pi} \left[ \tilde{z}_k \tilde{z}_k^T + 1 \right], \varphi_3 = E_{\pi} \left[ \tilde{z}_k c(x_k, u_k) \right].$$

For known $P$, $F(P)$ can be estimated using least squares from the collected data

$$\text{svec}(\tilde{F}(P)) = \Phi_M^T \Psi_M \text{svec}(P) + \Phi_M^T \Xi_M,$$
where \( M \in \mathbb{Z}_+, M > 0 \) and
\[
\Phi_M = \frac{1}{M} \sum_{k=0}^{M-1} \zeta_k \zeta_k^T, \quad \Psi_M = \frac{1}{M} \sum_{k=0}^{M-1} \zeta_k \zeta_{k+1}^T,
\]
\[
\Xi_M = \frac{1}{M} \sum_{k=0}^{M-1} \zeta_k c(x_k, u_k).
\]
In this way, (19) can be solved approximately and directly from the data by noticing that \( Q(P) = \mathcal{H}(F(P), 0) \). The O-LSPI is presented in Algorithm 1. Note that the same data matrices \( \Phi_M, \Psi_M \) and \( \Xi_M \) are reused for all iterations, thus O-LSPI is off-policy. The convergence of O-LSPI is proved in the following theorem.

**Theorem 3.** In Algorithm 1 under Assumption 2 for any initial stabilizing control gain \( \hat{K}_1 \) and any \( \epsilon > 0 \), there exist \( T_0 \in \mathbb{Z}_+ \) and \( M_0 \in \mathbb{Z}_+ \), such that for any \( T \geq T_0 \) and \( M \geq M_0 \), almost surely,
\[
\lim_{N \to \infty} \| \hat{P}_N - P^* \|_F < \epsilon
\]
and \( \hat{K}_i \) is stabilizing for all \( i = 1, \ldots, N \), where \( \hat{P}_N \) is the unique solution of (5) for \( K_i \).

The proof of Theorem 3 can be found in Appendix C. Let \( J_{O-LSPI}(\hat{K}_i; x) = \text{tr}(C^T \hat{P}_i C) \) denote the true cost induced by \( \hat{K}_i \). By Corollary 1 the task is to prove that there exist \( T_0 \in \mathbb{Z}_+ \) and \( M_0 \in \mathbb{Z}_+ \), such that for any \( T \geq T_0 \) and \( M \geq M_0 \), almost surely, \( \| \Delta G \|_F \leq \delta_2 \). For Algorithm 1
\[
\hat{G}_i = \hat{Q}_i, T - \left[ \begin{array}{cc} \hat{P}_i, T & 0 \\ 0 & 0 \end{array} \right].
\]
Using (21), we have
\[
\| \Delta G_i \|_F = \| \hat{G}_i - G_i \|_F \\
\leq \| \hat{Q}_i, T - Q(\hat{P}_i, T) \|_F + \| Q(\hat{P}_i, T) - Q(\hat{P}_i) \|_F + \| \hat{P}_i, T - \hat{P}_i \|_F.
\]
Since \( \hat{K}_1 \) is stabilizing, by the Birkhoff ergodic theorem 16.14], almost surely
\[
\lim_{M \to \infty} \Phi_M = \varphi_1, \quad \lim_{M \to \infty} \Psi_M = \varphi_2,
\]
\[
\lim_{M \to \infty} \Xi_M = \varphi_3.
\]
Using (24), Assumption 1 (18) and (22), we are able to show that there exist \( T_0 \) and \( M_0 \), independent of iteration index \( i \), such that for any \( T \geq T_0 \) and \( M \geq M_0 \), almost surely every term in (23) is less than \( \delta_2/3 \). Then Corollary 1 completes the proof.

**Algorithm 1: O-LSPI**

**Input:** Initial stabilizing controller \( \hat{K}_1 \), Number of policy iterations \( N \), Number of iteration for policy evaluation \( T \), Number of rollout \( M \), Exploration variance \( \sigma^2_n \).

1. Collect data with input \( u_k = -\hat{K}_1 x_k + v_k, v_k \sim N(0, \sigma^2_n I_m) \), to construct \( \Phi_M, \Psi_M \) and \( \Xi_M \).
2. for \( i = 1, \ldots, N - 1 \)
   3. \( \hat{P}_{i,0} \leftarrow 0; \)
   4. for \( j = 0, \ldots, T - 1 \)
      5. \( \text{svec}(\hat{F}_{i,j}) \leftarrow \Phi_M \hat{P}_i \text{svec}(\hat{P}_i) + \Phi_M \Xi_M; \)
      6. \( \hat{Q}_{i,j} \leftarrow \mathcal{H}(\hat{F}_{i,j}, 0); \)
      7. \( \hat{P}_{i,j+1} \leftarrow \mathcal{H}(\hat{Q}_{i,j}, \hat{K}_i); \)
   8. end
   9. \( \text{svec}(\hat{F}_{i,T}) \leftarrow \Phi_M \text{svec}(\hat{P}_i) + \Phi_M \Xi_M; \)
   10. \( \hat{Q}_{i,T} \leftarrow \mathcal{H}(\hat{F}_{i,T}, 0); \)
   11. \( \hat{K}_{i+1} \leftarrow [\hat{Q}_{i,T}]_u^{-1} [\hat{Q}_{i,T}]_{ux}; \)
12. end
13. return \( \hat{K}_N \).

**V. EXPERIMENTS**

We apply O-LSPI to the LQR problem studied in (22) with
\[
A = \left[ \begin{array}{ccc} 0.95 & 0.01 & 0 \\ 0.01 & 0.95 & 0.01 \\ 0 & 0.01 & 0.95 \end{array} \right], \quad B = \left[ \begin{array}{cc} 1 & 0.1 \\ 0 & 0.1 \\ 0 & 0 \end{array} \right],
\]
\[
C = S = I_3, \quad R = I_2.
\]
Note that in this example \( A \) is stable, so we just choose the initial stabilizing control gain to be \( \hat{K}_1 = 2 \alpha x_3 \). The exploration variance is set to \( \sigma^2_n = 1 \). All the experiments are conducted using MATLAB 2017b, on the NYU High Performance Computing Cluster Prince, with 4 CPUs and 16GB Memory. Algorithm 1 is implemented with increasing values of parameters \( N, T \) and \( M \), until the performance of the resulting control gain (almost) does not improve. This yields

\[ N = 5, T = 45 \text{ and } M = 10^6. \]

To investigate the performance of the algorithm with different values of \( M \) and \( T \), we conducted two sets of experiments: (a) Fix \( N = 5 \) and \( T = 45 \), and implement Algorithm 1 with increasing values of \( M \) from 200 to \( 10^6 \); (b) Fix \( N = 5 \) and \( M = 10^6 \), and implement Algorithm 1 with increasing values of \( T \) from 2 to 45. To evaluate the stability, we run Algorithm 1 for 100 times per set of parameters, and compute the fraction of times it produces stable policies in all phases (left column in Figure 1). To evaluate the optimality, the relative error of the cost function \( \text{tr}(C^T (\hat{P}_N - P^*) C) / \text{tr}(C^T P^* C) \) is calculated. The relative errors of 100 stable implementations of Algorithm 1 are collected (i.e., implementation that yields stabilizing control gains in all phases), based on which the sample average (middle column in Figure 1) and sample variance (right column in Figure 1) of the relative error are plotted.

In Figure 1 as the number of rollout \( M \) increases, the fraction of stability becomes one, and both the sample average and sample variance of relative error converge to zero. The fraction of stability is not sensitive to the number of iteration for policy evaluation \( T \). But as \( T \) increases, the sample average and sample variance of relative error improve and converge to zeros. These observations are consistent with our Theorem 3 thus are also consistent with our robustness analysis for policy iteration, since Theorem 3 is based on Corollary 1.

For comparison, the off-policy least-squares policy iteration algorithm LSPIv1 in (22) is also implemented, using the same setting with the first set of experiments of various \( M \) (upper row in Figure 1). The O-LSPI and LSPIv1 have similar performance for \( M \geq 10^4 \), while the performance of LSPIv1 is slightly better than that of O-LSPI for \( M < 10^4 \). This may be explained by the fact that the LSPIv1 in (22) assume knowledge of the matrix \( C \) in (8), which is not required in O-LSPI.

**VI. RELATED WORK**

Investigations of the robustness of policy iteration for problems with continuous state/control spaces are available in previous literature. In [9] Proposition 3.6], for discounted optimal control problems of discrete-time systems, it is reported that
\[
\lim_{i \to \infty} \max_{x \in K, u \in \mathcal{U}} (J^i(x) - J^*(x)) \leq \epsilon + 2\alpha \delta \left(1 - \alpha \right)^i,
\]
where \( J^i \) is the policy generated in \( i \)th iteration, \( \delta \) and \( \epsilon \) are the upper bounds of the errors in policy evaluation and policy improvement.
respectively, $0 < \alpha < 1$ is the discount factor. Bound (25) and our bound in Corollary 1 have the similar styles. However, in our setting the discount factor is $\alpha = 1$ so our bound cannot be implied by (25).

Utilizing the fact that Riccati operator is contractive in Thompson part metric [41], it is shown in [22] Appendix B that the convergence to the optimal solutions is still achieved in Thompson part metric, if the errors converge to zero. But it is unclear if this result could imply Theorem 2 and Corollary 1 in this paper. Sufficient conditions on the errors are given in [17] Chapter 2 and [9] Chapter 2 for continuous-time linear and nonlinear system dynamics respectively, to guarantee that the newly generated control policy is stabilizing and improved. The robustness analysis in this paper is parallel to that in [14]. However, since we are dealing with discrete-time systems here, the derivations and proofs are inevitably distinct.

In recent years, there have been resurgence research interests in LQR problems, about learning the optimal solutions from the input/state/output data. The model-based certainty equivalence methods explicitly estimate the values of $A$, $B$ and $C$ in (8) from data, and obtain near-optimal solutions based on the estimations, see [2], [33], [13], [38], [45], [12], [5], to name a few. The model-free methods aim at finding the near-optimal solutions directly from the data, without the estimations of system dynamics. Action-value model-free methods learn the value functions of policies, and then generate new (improved) policies based on the estimated value functions, see [10], [43], [22], [11], [44], [8]. Policy-gradient model-free methods directly learn the policies based on the gradient of some scalar performance measure with respect to the policy parameter, see [14], [11], [35], [31], [47], [56]. Derivative-free model-free methods randomly search in the parameter space of policies for the near-optimal solutions, without explicitly estimate the gradient, see [30], [29], [25]. Most of the model-free methods for LQR mentioned above are on-policy, fewer theoretical results exist for off-policy methods. Among the off-policy action-value model-free methods for LQR, the most related to our proposed O-LSPI are the LSPi1 in [22] and the MFLQ1 in [11]. However, (a) no convergence result is reported for LSPi1 in [22], and (b) MFLQ1 in [11] needs to learn the $P_K$ in (5) first in on-policy fashion, before it can learn the $Q(P_K)$ in (19) in off-policy fashion in each iteration, and (c) both the LSPi1 and MFLQ1 need the knowledge of matrix $C$ in [8], and (d) both the LSPi1 and MFLQ1 need to solve a pseudo-inverse problem in each iteration. In contrast, in our O-LSPI, (a) a convergence result is given (Theorem 3), and (b) both the $P_K$ and $Q(P_K)$ are learned in off-policy fashion (Lines 5 to 7 in Algorithm 1), and (c) no knowledge of $C$ is required, and (d) pseudo-inverse problem only needs to be solved once.

VII. Concluding Remarks

This paper analyzes the robustness of policy iteration for discrete-time LQR. It is proved that starting from any stabilizing initial policy, the solutions generated by policy iteration with errors are bounded and ultimately enter and stay in a neighbourhood of the optimal solution, as long as the errors are small and bounded. This result is employed to prove the convergence of the optimistic least-squares policy iteration (O-LSPI), a novel off-policy model-free RL algorithm for discrete-time LQR with additive stochastic noises in the dynamics. The theoretical results are verified by the experiments on a numerical example.

References

[1] Y. Abbasi-Yadkori, N. Lazic, and C. Szepesvari. Model-free linear quadratic control via reduction to expert prediction. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2019.
[2] Y. Abbasi-Yadkori and C. Szepesvri. Regret bounds for the adaptive control of linear quadratic systems. In Annual Conference on Learning Theory (COLT), 2011.
[3] M. Abeille and A. Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In International Conference on Machine Learning (ICML), 2018.
[4] R. P. Agarwal. Difference Equations and Inequalities: Theory, Methods, and Applications. Marcel Dekker, Inc, New York, 2nd edition, 2000.
[5] M. Bae, X. Guo, and A. Hu. Linear quadratic reinforcement learning: Sublinear regret in the episodic continuous-time framework. arXiv preprint arXiv:2006.15316, 2020.
[6] D. P. Bertsekas. Dynamic Programming and Optimal Control, volume 2. Athena Scientific, Belmont, MA, 1995.
[7] D. P. Bertsekas. Approximate policy iteration: A survey and some new methods. Journal of Control Theory and Applications, 2011.
[8] T. Bian and Z. P. Jiang. Continuous-time robust dynamic programming. SIAM Journal on Control and Optimization, 2019.
[9] C. I. Boussios. An approach for nonlinear control design via approximate dynamic programming. PhD thesis, Massachusetts Institute of Technology, 1998.
[10] S. J. Bradtke, B. E. Ydstie, and A. G. Barto. Adaptive linear quadratic control using policy iteration. In the Proceedings of Annual American Control Conference (ACC), 1994.
**APPENDIX A**

**USEFUL AUXILIARY RESULTS**

Some useful properties of $\mathcal{L}_X(\cdot)$ are provided below.

**Lemma 2.** Suppose $X \in \mathbb{R}^{n \times n}$ is stable.

(i) For each $Z \in \mathbb{R}^{n \times n}$, if $Y \in \mathbb{S}^n$ is the solution of equation $\mathcal{L}_X(Y) = -Z$, then

$$Y = \sum_{k=0}^{\infty} (X^T)^k ZX^k.$$

(ii) $\mathcal{L}_X(Y_1) \leq \mathcal{L}_X(Y_2) \implies Y_1 \geq Y_2$, where $Y_1, Y_2 \in \mathbb{S}^n$.

**Proof.**

(i) can be found in [15]. For (ii), let $x$ be the solution of system $x_{k+1} = Xx_k, x(0) = x_0$. Then,

$$x^T_0 Y_1 x_0 = \sum_{k=0}^{\infty} x^T_k (-\mathcal{L}_X(Y_1)) x_k \geq \sum_{k=0}^{\infty} x^T_k (-\mathcal{L}_X(Y_2)) x_k = x^T_0 Y_2 x_0.$$

$x_0$ is arbitrary, thus $Y_1 \geq Y_2$. 

When Definition [1] is truncated to the linear dynamical system [1], since $x^+ = 0$, we have the following definition.

**Definition 4.** System $x_{k+1} = O x_k$ is exponentially stable if for some $a > 0$ and $0 < b < 1$,

$$\|O^k\|_2 \leq ab^k.$$

Definition 4 is equivalent to the $(\tau, \rho)$-stability defined in [22, Definition 1], and is closely related to the strong stability defined in [12, Definition 5].

**Lemma 3 ([1] Remark 5.5.3.).** System $x_{k+1} = O x_k$ is stable if and only if it is exponentially stable.

Generally in Definition 4 $a$ and $b$ depend on the choice of matrix $O$. The next lemma shows that a set of (exponentially) stable systems can share the same constants $a$ and $b$.

**Lemma 4.** Let $\mathcal{O}$ be a compact set of stable matrices, then there exist an $a_0 > 0$ and $a_0 < b_0 < 1$, such that

$$\|O^k\|_2 \leq a_0 b_0^k, \quad \forall k \in \mathbb{Z}_+$$

for any $O \in \mathcal{O}$.

**Proof.** For each $O \in \mathcal{O}$, by [44, Lemma B.1] there exist $r > 0$, $a > 0$ and $0 < b < 1$, such that

$$\|O^k\|_2 \leq ab^k, \quad \forall k \in \mathbb{Z}_+$$

for all $\|O' - O\|_2 < r$. Then the compactness of $\mathcal{O}$ completes the proof.

The following lemma provides the relationship between operations vec(·) and svec(·).

**Lemma 5 ([28 Page 57]).** For $X \in \mathbb{S}^n$, there exists a unique matrix $D_n \in \mathbb{R}^{n \times \frac{n^2}{2} (n+1)}$ with full column rank, such that

$$\text{vec}(X) = D_n \text{svec}(X), \quad \text{svec}(X) = D_n^T \text{vec}(X).$$

$D_n$ is called the duplication matrix.

For $X \in \mathbb{R}^{n \times n}$, $Y \in \mathbb{R}^{m \times m}$, $X + \Delta X \in \mathbb{R}^{n \times n}$, $Y + \Delta Y \in \mathbb{R}^{m \times m}$, supposing $X$ and $X + \Delta X$ are invertible, the following inequality is repeatedly used

$$\|X^{-1}Y - (X + \Delta X)^{-1}(Y + \Delta Y)\|_F \leq \|X^{-1}Y - X^{-1}(Y + \Delta Y) + X^{-1}(Y + \Delta Y) - (X + \Delta X)^{-1}(Y + \Delta Y)\|_F \leq \|X^{-1}\|_F \|\Delta Y\|_F + \|X^{-1}\|_F \|X + \Delta X\|^{-1}_F \|\Delta Y\|_F \|\Delta X\|_F$$

\[\text{(26)}\]

**APPENDIX B**

**PROOF OF LEMMA 1**

Since $\mathcal{A}(\mathcal{K}(P^*))$ is stable, by continuity there always exists a $\delta_0 > 0$, such that $\mathcal{H}(P_i)$ is invertible, $\mathcal{A}(\mathcal{K}(P_i))$ is stable for all $P_i \in \mathcal{B}_{\delta_0}(P^*)$. Suppose $P_i \in \mathcal{B}_{\delta_0}(P^*)$. Subtracting

$$K_{i+1}^T B^T P^* A + A^T P^* B K_{i+1} - K_{i+1}^T \mathcal{H}(P^*) K_{i+1}$$

from both sides of the ARE (2) yields

$$\mathcal{L}_{A(\mathcal{K}(P_i))}(P^*) = -S - \mathcal{K}(P_i) R \mathcal{K}(P_i) + (\mathcal{K}(P_i) - \mathcal{K}(P^*))^T \mathcal{H}(P^*) (\mathcal{K}(P_i) - \mathcal{K}(P^*))$$

\[\text{(27)}\]
Subtracting (27) from (13), we have

\[
P_{k+1} - P^* = \mathcal{L}^{-1}_{\mathcal{A}(\mathcal{K}(P_{k+1}^i))} \left( \left( (\mathcal{K}(P_{k+1}) - \mathcal{K}(P^*))^T \mathcal{A}(P^*) \mathcal{K}(P_{k+1}) - \mathcal{K}(P^*) \right) \right).
\]

Taking norm on both sides of above equation, (12) yields

\[
\|P_{k+1} - P^*\|_F \leq \|\mathcal{A}(\mathcal{K}(P_{k+1}))^{-1}\|_2 \|\mathcal{A}(P^*)\|_F \|\mathcal{K}(P_{k+1}) - \mathcal{K}(P^*)\|_F.
\]

Since \(\mathcal{K}(\cdot)\) is locally Lipschitz continuous at \(P^*\), by continuity of matrix norm and matrix inverse, there exists a \(c_1 > 0\), such that

\[
\|P_{k+1} - P^*\|_F \leq c_1\|P_k - P^*\|_F, \quad \forall P_k \in \mathcal{B}(P^*).
\]

So for any \(0 < \sigma < 1\), there exists a \(\delta_0 \geq \delta_0 > 0\) with \(c_1\delta_0 \leq \sigma\). This completes the proof.

**APPENDIX C**

**PROOF OF THEOREM 2**

Before proving Theorem 2, we firstly prove some auxiliary lemmas. The Procedure 2 will exhibit a singularity, if \([\mathcal{G}_i]_{uu}\) in (11) is singular, or the cost \(\delta_1\) of \(K_{i+1}\) is infinity. The following lemma shows that if \(\Delta G_i\) is small, no singularity will occur. Let \(\delta_0\) be the one defined in the proof of Lemma 1 then \(\delta_0 \leq \delta_0\).

**Lemma 6.** For any \(\hat{P}_i \in \mathcal{B}(P^*)\), there exists a \(d(\delta_0) > 0\), independent of \(\hat{P}_i\), such that \(\hat{K}_{i+1}\) is stabilizing and \([\hat{G}_i]_{uu}\) is invertible, if \(\|\Delta G_i\|_F \leq 2\).

**Proof.** Since \(\hat{B}(P^*)\) is compact and \(\mathcal{A}(\mathcal{K}(\cdot))\) is a continuous function, set

\[
\mathcal{S} = \{\mathcal{A}(\mathcal{K}(\hat{P}_i)) | \hat{P}_i \in \mathcal{B}(P^*)\}
\]

is also compact. By continuity, for each \(X \in \mathcal{S}\), there exists a \(r(X) > 0\) such that any \(Y \in \mathcal{B}(X)\) is stable. The compactness of \(\mathcal{S}\) implies the existence of a \(r > 0\), such that each \(Y \in \mathcal{B}(X)\) is stable. Similarly, there exists \(d_1 > 0\) such that \([\hat{G}_i]_{uu}\) is invertible for all \(\hat{P}_i \in \mathcal{B}(P^*)\), if \(\|\Delta G_i\|_F \leq d_1\). Note that in policy improvement step of Procedure 2 (the policy update step in Procedure 2), the improved policy \(\hat{K}_{i+1} = \frac{1}{1+\|\hat{G}_i\|_2} \hat{G}_i^* \mathcal{G}_i^2 \hat{G}_i^*\mathcal{G}_i\) is continuous function of \(\mathcal{G}_i\), and there exists a \(0 < d_2 \leq d_1\), such that \(\mathcal{A}(\hat{K}_{i+1}) \in \mathcal{B}(\mathcal{A}(\mathcal{K}(\hat{P}_i)))\) for all \(\hat{P}_i \in \mathcal{B}(P^*)\), if \(\|\Delta G_i\|_F \leq d_2\). Thus \(\hat{K}_{i+1}\) is stabilizing. Setting \(d = d_2\) completes the proof.

By Lemma 8 if \(\|\Delta G_i\|_F \leq d\), the sequence \(\{\hat{P}_i\}_{i=0}^\infty\) satisfies (15). For simplicity, we denote \(\mathcal{E}(\hat{G}_i, \Delta G_i)\) in (15) by \(\mathcal{E}_i\). The following lemma gives an upper bound on \(\|\mathcal{E}_i\|_F\) in terms of \(\|\Delta G_i\|_F\).

**Lemma 7.** For any \(\hat{P}_i \in \mathcal{B}(P^*)\) and any \(c_2 > 0\), there exists a \(0 < \delta_1(\delta_0, c_2) \leq d\), independent of \(\hat{P}_i\), where \(d\) is defined in Lemma 6 such that

\[
\|\mathcal{E}_i\|_F \leq c_2\|\Delta G_i\|_F < c_2,
\]

if \(\|\Delta G_i\|_F < \delta_1\), where \(c_2(\delta_0) > 0\).

**Proof.** For any \(\hat{P}_i \in \mathcal{B}(P^*)\), \(\|\Delta G_i\|_F \leq d\), we have from (29)

\[
\|\mathcal{K}(\hat{P}_i) - \hat{K}_{i+1}\|_F \leq \|\hat{G}_i\|_2 \|\Delta G_i\|_F + \|\hat{G}_i\|_2 \|\Delta G_i\|_F \|\hat{G}_i\|_2 \|\Delta G_i\|_F \leq c_2(d_0, d)\|\Delta G_i\|_F,
\]

where the last inequality comes from the continuity of matrix inverse and the extremum value theorem. Define

\[
\hat{P}_i = \mathcal{L}^{-1}_{\mathcal{A}(\hat{K}_{i+1})} \left( -S - \hat{K}_{i+1}^T R \hat{K}_{i+1} \right), \quad \hat{P}_i = \mathcal{L}^{-1}_{\mathcal{A}(\mathcal{K}(\hat{P}_i))} \left( -S - \mathcal{K}(\hat{P}_i)^T R \mathcal{K}(\hat{P}_i) \right).
\]

Then by (12) and (15),

\[
\|\mathcal{E}_i\|_F = \|\text{vec}(\hat{P}_i - \hat{P}_i)\|_2
\]

\[
\text{vec}(\hat{P}_i) = \mathcal{A}^{-1}(\hat{K}_{i+1}) \text{vec}(S - \hat{K}_{i+1}^T R \hat{K}_{i+1})
\]

\[
\text{vec}(\hat{P}_i) = \mathcal{A}^{-1}(\mathcal{K}(\hat{P}_i)) \text{vec}(S - \mathcal{K}(\hat{P}_i)^T R \mathcal{K}(\hat{P}_i))
\]

Define

\[
\Delta \mathcal{A}_i = \mathcal{A}(\mathcal{K}(\hat{P}_i)) - \mathcal{A}(\hat{K}_{i+1}) \quad \Delta \mathcal{B}_i = \text{vec}(\mathcal{K}(\hat{P}_i)^T R \mathcal{K}(\hat{P}_i) - \hat{K}_{i+1}^T R \hat{K}_{i+1})
\]

Using (28), it is easy to check that \(\|\Delta \mathcal{A}_i\|_F \leq c_2\|\Delta G_i\|_F\), \(\|\Delta \mathcal{B}_i\|_2 \leq c_2\|\Delta G_i\|_F\), for some \(c_2(\delta_0, d) > 0\), \(c_2(\delta_0, d) > 0\). Then by (26)

\[
\|\mathcal{E}_i\|_F \leq \|\mathcal{A}^{-1}(\hat{K}_{i+1})\|\|\mathcal{A}^{-1}(\mathcal{K}(\hat{P}_i))\|_F \|\Delta G_i\|_F
\]

\[
\times \|S + \mathcal{K}(\hat{P}_i)^T R \mathcal{K}(\hat{P}_i)\|_F \|\Delta G_i\|_F
\]

\[
\leq c_2(\delta_0)\|\Delta G_i\|_F
\]

where the last inequality comes from the continuity of matrix inverse and Lemma 6. Choosing \(0 < \delta_1 \leq d\) such that \(c_2\delta_1^2 < c_2\) completes the proof.
Now we are ready to prove Theorem 2.

Proof of Theorem 2. Let \( c_2 = (1 - \sigma)\delta_0 \) in Lemma 7 and \( \delta_1 \) be equal to the \( \delta_1 \) associated with \( c_2 \). For any \( i \in \mathbb{Z}_+ \), if \( \hat{P}_i \in B_{\delta_0}(P^*) \), then \( [\hat{G}_i]_{uu} \) is invertible, \( \hat{K}_i \) is stabilizing and

\[
\| \hat{P}_{i+1} - P^* \|_F \leq \|E_i\|_F + \left\| \sum_{i \in \mathbb{Z}_+} \gamma A_i \right\|_F \| \hat{P}_{i+1} - P^* \|_F \\
\leq \frac{\sigma}{1 - \sigma} \| \hat{P}_i - P^* \|_F + c_3 \| \Delta G_i \|_\infty \\
\leq \frac{\sigma}{1 - \sigma} \| \hat{P}_i - P^* \|_F + c_3 \| \Delta G \|_\infty \\
< \sigma \delta_0 + c_3 \delta_1 < \sigma \delta_0 + c_2 = \delta_0,
\]  

(29) for all \( i \in \mathbb{Z}_+ \), thus by induction, (29) to (31) hold for all \( i \in \mathbb{Z}_+ \).

In terms of (iv) in Theorem 2, for any \( \epsilon > 0 \), there exists an \( i_0 \in \mathbb{Z}_+ \), such that \( \sup_{i} \| \Delta G_i \|_F \leq \gamma^{-1}(\epsilon/2) \). Take \( i_0 \geq i_1 \). For \( i \geq i_0 \), we have by (iv) in Theorem 2

\[
\| \hat{P}_i - P^* \|_F \leq \frac{\beta}{1 - \sigma} \| \hat{P}_i - P^* \|_F + \epsilon/2 \leq \beta(\epsilon/2) + \epsilon/2.
\]

where the second inequality is due to the boundedness of \( \hat{P}_i \). Since \( \lim_{i \to \infty} \beta(\epsilon/2) = 0 \), there is a \( i_3 \geq i_2 \) such that \( \beta(\epsilon/2) < \epsilon/2 \) for all \( i \geq i_3 \), which completes the proof. 

APPENDIX D

PROOF OF COROLLARY 1

Notice that all the conclusions of Corollary 1 can be implied by Theorem 2 if

\[
\delta_2 < \min(\gamma^{-1}(\epsilon), \delta_1), \quad \hat{P}_i \in B_{\delta_0}(P^*)
\]

for Procedure 2. Thus the proof of Corollary 1 reduces to the proof of the following lemma.

Lemma 8. Given a stabilizing \( \hat{K}_1 \), there exist \( \delta_2 < \min(\gamma^{-1}(\epsilon), \delta_1) \), \( i \in \mathbb{Z}_+ \), \( \Pi_2 > 0 \) and \( \kappa_2 > 0 \), such that \( [\hat{G}_i]_{uu} \) is invertible, \( \hat{K}_i \) is stabilizing, \( \| \hat{P}_i \|_F < \Pi_2 \), \( \| \hat{K}_i \|_F < \kappa_2 \), \( i = 1, \ldots, i \), \( \hat{P}_i \in B_{\delta_0}(P^*) \), as long as \( \| \Delta G \|_\infty < \delta_2 \).

The next two lemmas, inspired by [16] Lemma 5.1, state that under certain conditions on \( \| \Delta G \|_F \), each element in \{\( \hat{K}_i \)\}_{i=1}^{\infty} \) is stabilizing, each element in \{\( [\hat{G}_i]_{uu} \)\}_{i=1}^{\infty} \) is invertible and \{\( \hat{P}_i \)\}_{i=1}^{\infty} \) is bounded. For simplicity, in the following we assume \( S > I_n \) and \( R > I_m \). All the proofs still work for any \( S > 0 \) and \( R > 0 \), by suitable rescaling.

Lemma 9. If \( \hat{K}_i \) is stabilizing, then \( [\hat{G}_i]_{uu} \) is nonsingular and \( \hat{K}_i+1 \) is stabilizing, as long as \( \| \Delta G_i \|_F < a_\epsilon \), where

\[
a_\epsilon = \left( m(\sqrt{n} + \| \hat{K}_i \|_F)^2 + m(\sqrt{n} + \| \hat{K}_{i+1} \|_F^2)^2 \right)^{-1}.
\]

Furthermore,

\[
\| \hat{K}_{i+1} \|_F \leq 2 \| R^{-1} \|_F (1 + \| B^T \hat{P}_i A \|_F).
\]

Proof. By definition,

\[
\| [\hat{G}_i]_{uu} - [\hat{G}_i]_{uu} \|_F < a_\epsilon \| [\hat{G}_i]_{uu} \|_F.
\]

Since \( S > I_n \), the eigenvalues \( \lambda_j([\hat{G}_i]_{uu}) \in (0,1) \) for all \( 1 \leq j \leq m \). Then by the fact that for any \( X \in \mathbb{S}^n \)

\[
\| X \|_F = \| A X \|_F, \quad \lambda_X = \text{diag}\{\lambda_1(X), \ldots, \lambda_n(X)\},
\]

we have

\[
\| [\hat{G}_i]_{uu}^{-1} ([\hat{G}_i]_{uu} - [\hat{G}_i]_{uu}) \|_F < a_\epsilon \sqrt{n} < 0.5.
\]

(33)

Thus by [16] Section 5.8, \( [\hat{G}_i]_{uu} \) is invertible.

For any \( x \in \mathbb{R}^n \) on the unit ball, define

\[
\mathcal{X}_{\hat{K}_i} = \begin{bmatrix} I & \hat{K}_i \\ -\hat{K}_i^T & I \end{bmatrix} x x^T \begin{bmatrix} I & -\hat{K}_i^T \\ \hat{K}_i & I \end{bmatrix}.
\]

From (10) and (11) we have

\[
x^T \mathcal{H}(\hat{G}_i, \hat{K}_i)x = \text{tr}(\hat{G}_i \mathcal{X}_{\hat{K}_i}) = 0,
\]

and

\[
\text{tr}(\hat{G}_i \mathcal{X}_{\hat{K}_{i+1}}) = \min_{K \in \mathbb{R}^{m \times n}} \text{tr}(\hat{G}_i \mathcal{X}_K).
\]
Then
\[
\text{tr}(\hat{G}, X_{K_{i+1}^+}) \leq \text{tr}(\hat{G}, X_{K_i}) + \|\Delta G_t\| F \text{tr}(11^T X_{K_{i+1}^+}) \leq \text{tr}(\hat{G}, X_{K_i}) + \|\Delta G_t\| F 1^T (X_{K_i} + X_{K_{i+1}^+}) 1 \leq \|\Delta G_t\| F 1^T (X_{K_i} + X_{K_{i+1}^+}) 1,
\]
\[
(34)
\]
where \(X_{K_i} \) denotes the matrix obtained from \(X_{K_i} \) by taking the absolute value of each entry. Thus by (34) and the definition of \(\hat{G} \), we have
\[
x^T L_{A(\hat{K}_{i+1})} (\hat{P}) x + \epsilon_1 \leq 0
\]
where
\[
\epsilon_1 = x^T (S + \hat{K}^T_{i+1} R \hat{K}_{i+1}) x - \|\Delta G_t\| F 1^T (X_{K_i} + X_{K_{i+1}^+}) 1.
\]
For any \(x\) on the unit ball, \(|1^T x| \leq \sqrt{n}\). Similarly, for any \(K \in \mathbb{R}^{m \times n}\), by the definition of induced matrix norm, \(|1^T K x| \leq \|K\|_2 \sqrt{n}\). This implies
\[
\text{tr}(1^T ) \leq m(\sqrt{n} + \|K\|_2)^2. \quad \text{Thus}
\]
\[
\|\Delta G_t\| F 1^T (X_{K_i} + X_{K_{i+1}^+}) 1 < 1.
\]
Then \(S > I_n\) leads to
\[
x^T \left(A(\hat{K}_{i+1})^T \hat{P} A(\hat{K}_{i+1}) - \hat{P}\right) x < 0
\]
for all \(x\) on the unit ball. So \(\hat{K}_{i+1}\) is stabilizing by the Lyapunov criterion \cite[Lemma 12.2']{16}. By definition,
\[
\|\hat{K}_{i+1}\|_F \leq \|\hat{G}\|_F (1 + \|B^T \hat{P} A\|_F) \leq \|\hat{G}\|_F (1 - \|\hat{G}\|_F (\hat{G} - [\hat{G}])_u)\|_F^{-1} (1 + \|B^T \hat{P} A\|_F) \leq 2\|R^{-1}\|_F (1 + \|B^T \hat{P} A\|_F).
\]
(36)
where the second inequality comes from \cite[Inequality (5.8.2)]{16}, and the last inequality is due to (33). This completes the proof.

\[\square\]

Lemma 10. For any \(\tilde{i} \in \mathbb{Z}_+\), \(\tilde{i} > 0\), if
\[
\|\Delta G_t\|_F < (1 + \tilde{i}^2)^{-1} \alpha_i, \quad i = 1, \ldots, \tilde{i},
\]
(37)
where \(\alpha_i\) is defined in Lemma \[2\] then
\[
\|\hat{P}_i\|_F \leq 6\|\hat{P}_1\|_F, \quad \|\hat{K}_i\|_F \leq C_0,
\]
for \(i = 1, \ldots, \tilde{i}\), where
\[
C_0 = \max \left\{\|\hat{K}_1\|_F, 2\|R^{-1}\|_F \left(1 + 6\|B^T \|_F\|\hat{P}_1\|_F\|A\|_F\right)\right\}.
\]
Proof. Inequality (35) yields
\[
L_{A(\hat{K}_{i+1})} (\hat{P}_i) + (S + \hat{K}^T_{i+1} R \hat{K}_{i+1}) - \epsilon_2, i < 0
\]
(38)
where
\[
\epsilon_{2,i} = \|\Delta G_t\|_F 1^T (X_{K_i} + X_{K_{i+1}^+}) 1 < 1.
\]
(39)
Inserting (35) into above inequality, and using (31) in Lemma \[2\] we have
\[
\hat{P}_{i+1} = \hat{P}_i + \epsilon_{2,i} L^{-1}_{A(\hat{K}_{i+1})} (-I).
\]
(40)
With \(S > I_n\), (38) yields
\[
L_{A(\hat{K}_{i+1})} (\hat{P}_i) + (1 - \epsilon_{2,i}) I < 0.
\]
Similar to (39), we have
\[
L^{-1}_{A(\hat{K}_{i+1})} (-I) < \frac{1}{1 - \epsilon_{2,i}} \hat{P}_i.
\]
(41)
From (39) and (40), we obtain
\[
\hat{P}_{i+1} \leq \left(1 + \frac{\epsilon_{2,i}}{1 - \epsilon_{2,i}}\right) \hat{P}_i.
\]
(42)
By definition of \(\epsilon_{2,i}\) and condition (37),
\[
\frac{\epsilon_{2,i}}{1 - \epsilon_{2,i}} \leq \frac{1}{\tilde{i}^2}, \quad i = 1, \ldots, \tilde{i}.
\]
Then [20] Theorem 3] yields \( \tilde{P}_i \leq 6\tilde{P}_1, \quad i = 1, \ldots, \bar{i}. \)

An application of [32] completes the proof.

Now we are ready to prove Lemma 8.

**Proof of Lemma 8.** Consider Procedure 2 confined to the first \( \bar{i} \) iterations, where \( \bar{i} \) is a sufficiently large integer to be determined later in this proof. Suppose

\[
\|\Delta G_i\|_F < b_1 \leq \frac{1}{2m(1+t^2)} (\sqrt{n} + C_0)^{-2}.
\]  

(41)

Condition [41] implies condition [37]. Thus \( \hat{K}_i \) is stabilizing, \( [G_i]_{n \times n} \) is invertible, \( \|\hat{P}\|_F \) and \( \|\hat{K}_i\|_F \) are bounded. By [10] we have

\[
\mathcal{L}_{A(K_{i+1})}(\hat{P}_{i+1} - \hat{P}_i) = -S - \hat{K}_{i+1}^T R \hat{K}_{i+1} - \mathcal{L}_{A(K_{i+1})}(\hat{P}_i).
\]

Letting \( E_i = \hat{K}_{i+1} - \mathcal{H}(\hat{P}_i) \), the above equation can be rewritten as

\[
\hat{P}_{i+1} = \hat{P}_i - N(\hat{P}_i) + \mathcal{L}^{-1}_{A(\mathcal{H}(\hat{P}_i))}(E_i),
\]

(42)

where \( N(\hat{P}_i) = \mathcal{L}^{-1}_{A(\mathcal{H}(\hat{P}_i))} \circ \mathcal{R}(\hat{P}_i) \), and

\[
\mathcal{R}(Y) = A^T Y A - Y - A^T Y B (R + B^T Y B)^{-1} B^T Y A + S,
\]

\[
E_i = -E_i^T \mathcal{R}(\hat{P}_{i+1}) E_i + E_i^T \mathcal{R}(\hat{P}_{i+1}) \left( \mathcal{H}(\hat{P}_{i+1}) - \mathcal{H}(\hat{P}_i) \right) + \left( \mathcal{H}(\hat{P}_{i+1}) - \mathcal{H}(\hat{P}_i) \right)^T \mathcal{R}(\hat{P}_{i+1}) E_i.
\]

Given \( \hat{K}_1 \), let \( M_1 \) denote the set of all possible \( \hat{P}_1 \), generated by [42] under condition [41]. By definition, \( \{M_j\}_{j=1}^{\infty} \) is a nondecreasing sequence of sets, i.e., \( M_1 \subset M_2 \subset \ldots \). Define \( M = \bigcup_{j=1}^{\infty} M_j, D = \{P \in \mathbb{R}^n \mid \|P\|_F \leq 6\|\hat{P}_1\|_F \} \). Then by Lemma [10] and Theorem [1], \( M \subset D; M \) is compact; \( A(\mathcal{H}(P)) \) is stable for any \( P \in M \).

Now we prove that \( N(P^1) \) is Lipschitz continuous on \( M \). Using [12], we have

\[
\|N(P^1) - N(P^2)\|_F = \|\mathcal{H}^{-1}(A(\mathcal{H}(P^1))) \mathcal{R}(P^1) - \mathcal{H}^{-1}(A(\mathcal{H}(P^2))) \mathcal{R}(P^2)\|_2 \\
\leq \|\mathcal{H}^{-1}(A(\mathcal{H}(P^1)))\|_2 \|\mathcal{R}(P^1) - \mathcal{R}(P^2)\|_F + \|\mathcal{R}(P^2)\|_F \|\mathcal{H}^{-1}(A(\mathcal{H}(P^1))) - \mathcal{H}^{-1}(A(\mathcal{H}(P^2)))\|_2 \\
\leq L \|P^1 - P^2\|_F,
\]

(43)

where the last inequality is due to the fact that matrix inversion, \( A(\cdot), \mathcal{H}(\cdot) \) and \( \mathcal{R}(\cdot) \) are locally Lipschitz, thus Lipschitz on compact set \( M \) with some Lipschitz constant \( L > 0 \).

Define \( \{P_{k|i}\}_{k=0}^{\infty} \) as the sequence generated by [13] with \( P_{0|i} = \hat{P}_i \). Similar to [42], we have

\[
P_{k+1|i} = P_{k|i} - N(P_{k|i}), \quad k \in \mathbb{Z}_+.
\]

(44)

By Theorem [1] and the fact that \( M \) is compact, there exists \( \delta_0 \in \mathbb{Z}_+ \), such that

\[
\|P_{0|i} - P^*\|_F < \delta_0/2, \quad \forall P_{0|i} \in M.
\]

(45)

Suppose

\[
\|\mathcal{L}^{-1}_{A(\mathcal{H}(P_{i+k}))}(E_{i+k})\|_F < \mu, \quad k = 0, \ldots, \bar{i} - i.
\]

(46)

We find an upper bound on \( \|P_{k|i} - \hat{P}_{i+k}\|_F \). Notice that from [42] and [44],

\[
P_{k|i} = P_{0|i} - \sum_{j=0}^{k-1} N(P_{j|i}), \quad \hat{P}_{i+k} = \hat{P}_i - \sum_{j=0}^{k-1} N(\hat{P}_{i+j}) + \sum_{j=0}^{k-1} \mathcal{L}^{-1}_{A(\hat{P}_{i+j})}(E_{i+j}).
\]

Then [43] and [46] yield

\[
\|P_{k|i} - \hat{P}_{i+k}\|_F \leq \mu + \sum_{j=0}^{k-1} L \|P_{j|i} - \hat{P}_{i+j}\|_F.
\]

An application of the Gronwall inequality [4] Theorem 4.1.1.4) to the above inequality implies

\[
\|P_{k|i} - \hat{P}_{i+k}\|_F \leq \mu + \mu \sum_{j=0}^{k-1} j(1 + L)^{k-j-1}.
\]

(47)

By [1] in Lemma [2],

\[
\mathcal{L}^{-1}_{A(\mathcal{H}(P_{i+1}))}(-I) = \sum_{k=0}^{\infty} (A^T(\mathcal{H}(P^1)))^k(A(\mathcal{H}(P^1)))^k, \quad P^1 \in M.
\]
By definition of the set $\mathcal{M}$, there exist $G^1$ and stabilizing control gain $K^1$ associated with $P^1 \in \mathcal{M}$, such that $\mathcal{H}(G^1, K^1) = 0$. So for any $x \in \mathbb{R}^n$, 
\[
x^T R(P)x = \min_{K \in \mathbb{R}^{m \times n}} x^T \mathcal{H}(G^1, K)x \leq 0.
\]
This implies 
\[
\mathcal{L}_A(\mathcal{H}(P^1)) (P^1) \leq -S - \mathcal{K}^T \mathcal{H}(P^1) \mathcal{K} < -I.
\]
An application of \(41\) in Lemma \[2\] to the above inequality leads to 
\[
\|\mathcal{L}_A^{-1}(\mathcal{H}(P^1)) (-I)\|_F < P^1 \leq 6\|\hat{P}_1\|_F, \quad \forall P^1 \in \mathcal{M} \subset \mathcal{D}.
\] (48)
Then the error term in \(42\) satisfies 
\[
\|\mathcal{L}_A^{-1}(\mathcal{H}(P^1)) (\delta_{i1})\|_F = \left\| \sum_{k=0}^{\infty} (A^T(\mathcal{K}(\hat{P}_1)))^k \otimes (A(\mathcal{K}(\hat{P}_1)))^k \text{vec} (\delta_{i1}) \right\|_2 \leq C_1 \|\delta_{i1}\|_F,
\] (49)
where $C_1$ is a constant and the inequality is due to \(48\).

Let $\hat{i} > k_0$, and $k = \hat{k}_0$, $i = \hat{i} - k_0$ in \(47\). Then by condition \(41\), Lemma \[10\] \[46\], \[47\], and \[49\], there exists $i_0 \in \mathbb{Z}_+$, $i_0 > k_0$, such that $\|P_{k_0}\|_F < \delta_0/2$, for all $i \geq i_0$. Setting $i = \hat{i} - k_0$ in \(45\), the triangle inequality yields $\hat{P}_1 \in \mathcal{B}_{k_0}(P^\infty)$, for $\hat{i} \geq i_0$. Then in \[41\], choosing $i \geq i_0$ such that $\delta_2 = \delta_1 < \min(\gamma^{-1}(\epsilon), \delta_1)$ completes the proof. \square

### Appendix E

#### Proof of Theorem \[3\]

For given $\hat{K}_1$, let $\mathcal{F}$ denote the set of control gains (including $\hat{K}_1$) generated by Procedure \[2\] with all possible $\{\Delta G_i\}_{i=1}^{\infty}$ satisfying $\|\Delta G\|_\infty < \delta_2$, where $\delta_2$ is the one in Corollary \[1\] The following result is firstly derived.

**Lemma 11.** Under Assumption \[1\] there exist $T_0 \in \mathbb{Z}_+$ and $M_0 \in \mathbb{Z}_+$, such that for any $T \geq T_0$ and any $M \geq M_0$, $\hat{K}_i \in \mathcal{F}$ implies $\|\Delta G_i\|_F < \delta_2$, almost surely.

**Proof.** The task is to show that each term in \(23\) is less than $\delta_2/3$ almost surely.

We firstly study term $\|\hat{P}_i T - P^1\|_F$. Define $\hat{p}_{i,j} = \text{vec}(\hat{P}_{i,j})$, then by Lemma \[5\] lines from \[5\] to \[7\] in Algorithm \[1\] can be rewritten as 
\[
\hat{p}_{i,j+1} = \text{vec}(\mathcal{H}(\text{svec}^{-1}(\Phi_i^1 \Psi_{z,M} \text{svec}(\hat{P}_{i,j}) + \Phi_i^1 \Xi_{z,M}(0), \hat{K}_i))) = T^1(\Phi_{z,M}, \Psi_{z,M}, \hat{K}_i) \hat{p}_{i,j} + T^2(\Phi_{z,M}, \Xi_{z,M}, \hat{K}_i), \quad \hat{p}_{i,0} = 0,
\] (50)
where $Y = \text{svec}^{-1}(\text{svec}(Y))$, for any $Y \in \mathbb{S}^n$, and 
\[
T^1(\Phi_{z,M}, \Psi_{z,M}, \hat{K}_i) = \begin{pmatrix} I_n - \hat{K}_i^T \otimes I_n - \hat{K}_i^T \end{pmatrix} \begin{pmatrix} I_{m+n}, 0 \otimes I_{m+n}, 0 \end{pmatrix} D_{m+n+1}(\Phi_{z,M}^1 \Psi_{z,M} \hat{D}_i),
\]
\[
T^2(\Phi_{z,M}, \Xi_{z,M}, \hat{K}_i) = \begin{pmatrix} I_n - \hat{K}_i^T \otimes I_n - \hat{K}_i^T \end{pmatrix} \begin{pmatrix} I_{m+n}, 0 \otimes I_{m+n}, 0 \end{pmatrix} D_{m+n+1}(\Phi_{z,M}^1 \Xi_{z,M}).
\]
Set $K = \hat{K}_i$ and insert \(22\) into \(18\). Similar derivations yield 
\[
\hat{p}_{i,j+1} = T^1(\varphi_1, \varphi_2, \hat{K}_i) \hat{p}_{i,j} + T^2(\varphi_1, \varphi_3, \hat{K}_i), \quad \hat{p}_{i,0} = 0.
\] (51)
By Assumption \[1\] \(51\) is identical to 
\[
\hat{p}_{i,j+1} = (A - B \hat{K}_i) \hat{p}_{i,j} + \text{vec}(S + \hat{K}_i^T R \hat{K}_i), \quad \hat{p}_{i,0} = 0,
\] (52)
with 
\[
T^1(\varphi_1, \varphi_2, \hat{K}_i) = (A - B \hat{K}_i) \otimes (A - B \hat{K}_i)^T, \quad T^2(\varphi_1, \varphi_3, \hat{K}_i) = \text{vec}(S + \hat{K}_i^T R \hat{K}_i).
\] (53)
Since $\hat{K}_i \in \mathcal{F}$ is stabilizing, 
\[
\lim_{j \to \infty} \hat{P}_{i,j} = \hat{P}_i
\] (54)
where $\hat{p}_{i,j} = \text{vec}(\hat{P}_{i,j})$ and $\hat{P}_i$ is the unique solution of \(5\) with $K = \hat{K}_i$. By definition and Corollary \[1\] $\mathcal{F}$ is bounded, thus compact. Let $\mathcal{V}$ be the set of unique solutions of \(5\) with $K \in \mathcal{F}$. Then by Corollary \[1\] $\mathcal{V}$ is bounded. So any control gain in $\mathcal{F}$ is stabilizing, otherwise it contradicts the boundedness of $\mathcal{V}$. Define 
\[
\mathcal{F}_1 = \{(A - BK)^T \otimes (A - BK)^T | K \in \mathcal{F}\}.
\]
By continuity, $\mathcal{F}_1$ is a compact set of stable matrices, and there exists a $\delta_3 > 0$, such that any $X \in \mathcal{F}_2$ is stable, where 
\[
\mathcal{F}_2 = \{X | X \in \mathcal{B}_{\delta_3}(Y), Y \in \mathcal{F}_1\}.
\]
Define 
\[
\Delta T_{M,i} = T^1(\varphi_1, \varphi_2, \hat{K}_i) - T^1(\Phi_{z,M}, \Psi_{z,M}, \hat{K}_i), \quad \Delta T_{M,i}^2 = T^2(\varphi_1, \varphi_3, \hat{K}_i) - T^2(\Phi_{z,M}, \Xi_{z,M}, \hat{K}_i).
\]
The boundedness of $\mathcal{F}$, \(24\) and \(53\) imply the existence of $M_1 > 0$, such that for any $M \geq M_1$, any $\hat{K}_i \in \mathcal{F}$, almost surely 
\[
T^1(\Phi_{z,M}, \Psi_{z,M}, \hat{K}_i) \in \mathcal{F}_2, \quad T^2(\Phi_{z,M}, \Xi_{z,M}, \hat{K}_i) < C_2,
\] (55)
where $C_2 > 0$ is a constant. Then (50) admits a unique stable equilibrium, that is,
\[
\lim_{j \to \infty} \hat{P}_{i,j} = \hat{P}_i
\]
for some $\hat{P}_i \in S^n$, and from (50), (51), (54) and (56), we have
\[
\hat{p}_i = \text{vec}(\hat{P}_i) = \left( I_{n^2} - T^1(\varphi_1, \varphi_2, \hat{K}_i) \right)^{-1} T^2(\varphi_1, \varphi_3, \hat{K}_i),
\]
\[
\hat{p}_i = \text{vec}(\hat{P}_i) = \left( I_{n^2} - T^1(\Phi_i, \Psi_i, \hat{K}_i) \right)^{-1} T^2(\Phi_i, \Xi_i, \hat{K}_i).
\]
Thus by (26), for any $M \geq M_1$, any $\hat{K}_i \in F$, almost surely
\[
\|\hat{P}_i - \hat{P}_i\|_F \leq \left\| \left( I_{n^2} - T^1(\varphi_1, \varphi_2, \hat{K}_i) \right)^{-1} \right\|_F \left( \|\Delta T^2_{M,i}\|_2 + \left\| \left( I_{n^2} - T^1(\Phi_i, \Psi_i, \hat{K}_i) \right)^{-1} \right\|_F \left\| T^2(\Phi_i, \Xi_i, \hat{K}_i) \right\|_2 \|\Delta T^1_{M,i}\|_F \right)
\leq c_8\|\Delta T^2_{M,i}\|_F + c_9\|\Delta T^1_{M,i}\|_F
\]
where $c_8$ and $c_9$ are some positive constants, and the last inequality is due to (53), (55) and the fact that $F_1$ and $F_2$ are compact sets of stable matrices. Then for any $\epsilon_1 > 0$, the boundedness of $F$ and (24) implies the existence of $M_2 \geq M_1$, such that for any $M \geq M_2$, almost surely
\[
\|\hat{P}_i - \hat{P}_i\|_F < \epsilon_1/2,
\]
as long as $\hat{K}_i \in F$. By Lemma 4 and (55), for any $M \geq M_2$ and any $\hat{K}_i \in F$,
\[
\|\hat{p}_{i,j+1} - \hat{p}_i\|_2 \leq a_0 b_0^i \|\hat{p}_i\|_2 \leq a_1 b_0^i,
\]
for some $a_0 > 0$, $1 > b_0 > 0$ and $a_1 > 0$. Therefore there exists a $T_1 > 0$, such that for any $T \geq T_1$, and any $M \geq M_2$, almost surely
\[
\|\hat{P}_{i,T} - \hat{P}_i\|_F < \epsilon_1/2,
\]
as long as $\hat{K}_i \in F$. Synthesizing (57) and (58) yields
\[
\|\hat{P}_{i,T} - \hat{P}_i\|_F < \epsilon_1,
\]
almost surely for any $T \geq T_1$, any $M \geq M_2$, as long as $\hat{K}_i \in F$. Since $\epsilon_1$ is arbitrary, we can choose $\epsilon_1$ such that almost surely
\[
\|\hat{P}_{i,T} - \hat{P}_i\|_F < \epsilon_1 \leq \frac{\delta_2}{3}.
\]
Secondly, for term $\|Q(\hat{P}_{i,T}) - Q(\hat{P}_i)\|_F$, by (59) there exist $T_2 \geq T_1$, $M_3 \geq M_2$, such that almost surely
\[
\|Q(\hat{P}_{i,T}) - Q(\hat{P}_i)\|_F < \frac{\delta_3}{3}
\]
for any $T \geq T_2$, any $M \geq M_3$, as long as $\hat{K}_i \in F$.
Finally, since $V$ is bounded, by (59) $\hat{P}_{i,T}$ is also almost surely bounded. Thus from lines 9 to 10 in Algorithm 1 and (24), there exists $M_4 \geq M_5$, such that
\[
\|\hat{Q}_{i,T} - Q(\hat{P}_{i,T})\|_F < \frac{\delta_2}{3}
\]
for any $M \geq M_4$ and any $T \geq T_2$, as long as $\hat{K}_i \in F$.
Setting $T_0 = T_2$ and $M_0 = M_4$ yields $\|\Delta G_{1i}\| < \delta_2$.

Now we are ready to prove Theorem 3.

**Proof.** Since $\hat{K}_i \in F$, Lemma 11 implies $\|\Delta G_{1i}\|_F < \delta_2$ almost surely. By definition, $\hat{K}_2 \in F$. Thus $\|\Delta G_{1i}\|_F < \delta_2, i = 1, 2, \cdots$ almost surely by mathematical induction. Then Corollary 1 completes the proof. \[\square\]