SIAM J. Control Optim.
Vol. 57, No. 6, pp. 4150–4174

CONTINUOUS-TIME ROBUST DYNAMIC PROGRAMMING[*]

TAO BIAN[†] AND ZHONG-PING JIANG[‡]

Abstract. This paper presents a new theory, known as robust dynamic programming, for a class of continuous-time dynamical systems. Different from traditional dynamic programming (DP) methods, this new theory serves as a fundamental tool to analyze the robustness of DP algorithms, and, in particular, to develop novel adaptive optimal control and reinforcement learning methods. In order to demonstrate the potential of this new framework, two illustrative applications in the fields of stochastic and decentralized optimal control are presented. Two numerical examples arising from both finance and engineering industries are also given, along with several possible extensions of the proposed framework.

Key words. dynamic programming, stochastic optimal control, adaptive optimal control, robust control

AMS subject classifications. 68Q25, 68R10, 68U05

DOI. 10.1137/18M1214147

1. Introduction.

In 1952, Bellman proposed the original idea of dynamic programming (DP) [4] to solve a class of optimization problems subject to a controlled process that is usually described by a Markov decision process (MDP), a difference equation, or a differential equation. Over the past several decades, DP and its extensions [5, 20, 6, 7, 8] have attracted a significant amount of attention, because of the vital role they have played in several popular fields including reinforcement learning (RL) [50], finance [40], and biological control [53], to name a few. Depending on the form (discrete-time versus continuous-time) used to describe the dynamical system in question, DP problems can be solved by finding the solution to either the Bellman equation or the Hamilton–Jacobi–Bellman (HJB) equation. However, due to the complex nature of these equations, the optimal solution cannot be obtained analytically in most cases. Numerous methods, including policy iteration (PI) [20, 30, 3, 11] and value iteration (VI) [5, 9, 14, 13], have been developed to approximate the solutions of these equations. Unfortunately, these algorithms suffer from serious usage limitations, due to the limited information available and the presence of various types of disturbances in practical problems. Nevertheless, from a control theory point of view, we identify two perspectives to address these issues. The first, which we refer to as the "adaptive control perspective," aims at learning the unknown components in DP algorithms directly from available online/offline data. Based on the problem formulation, such unknown components can be the Q-factor [54], the policy gradient [50, Chapter 13], and the value function [55]. Indeed, the majority of existing adaptive optimal control and DP methods [10, 6, 7, 35, 24, 57] fall into this category, and RL is also considered as a machine learning reinterpretation of direct adaptive control [51]. The main advantage of these methods is that they are effective in tackling the presence of

[*] Received by the editors September 17, 2018; accepted for publication (in revised form) September 11, 2019; published electronically December 12, 2019. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views and policies of Bank of America Merrill Lynch.
https://doi.org/10.1137/18M1214147

Funding: The work of the second author was partially supported by the National Science Foundation under grants ECCS-1230040, ECCS-1501044, and EPCN-1903781.

[†] Bank of America Merrill Lynch, One Bryant Park, New York, NY 10036 (tbian@nyu.edu).

[‡] Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, 6 Metrotech Center, Brooklyn, NY 11201 (zjiang@nyu.edu).

static uncertainties such as the unknown parameters in the DP algorithm. As a result, this allows the DP problem to be solved without directly using the knowledge of the underlying system (also known as the environment in RL), i.e., the optimal solution is obtained in a model-free manner. In spite of its popularity, the adaptive control perspective is not effective in tackling the presence of dynamic uncertainties [38] in DP algorithms. Such dynamic uncertainty may be caused by coupling the standard DP algorithm with other numerical algorithms, where each of these algorithms then serves as a dynamic uncertainty in this hybrid algorithm. It may also arise from the decentralized DP problem, where each node in a large-scale network executes its own version of the DP algorithm and interacts with its neighbors through the outputs and inputs. In this case, the algorithm executed in its neighboring nodes can be considered as dynamic uncertainty to the node itself. Existing learning-based DP algorithms are not directly applicable to handle this type of disturbance.

The second perspective, which we refer to as the "robust control perspective," aims at strengthening the DP algorithm so that it is robust to the presence of disturbance. A remarkable feature of this type of method is that it is effective in dealing with both static and dynamic uncertainties. However, unlike the adaptive control perspective, the development in this direction is still rudimentary. Only a few results [22, 39, 37] have been developed along this track to solve DP and RL problems, in which case the authors only considered the static uncertainty caused by the unknown transition probability measures. As a result, how to develop DP algorithms that are robust to both static and dynamic uncertainties remains an open problem. In addition, existing robust DP methods are only available for MDPs. In particular, there is no robust DP solution for dynamical systems described by differential equations.

In this paper, we propose a novel robust DP theory for continuous-time linear dynamical systems. Compared with traditional DP and adaptive optimal control theory, we take a completely different path to investigate DP methods from a viewpoint of nonlinear system theory [28] and small-gain theory [60, 27]. As a consequence, we provide a complete robustness analysis on the DP algorithm, under multiple types of uncertainties, including external disturbance, dynamic uncertainty, and stochastic noise, that cannot be dealt with directly by previously known results. The proposed robust DP framework is based on the dynamic property of the differential matrix Riccati equation (DMRE). Recall from [59, 32] that under observability and stabilizability assumptions, the unique symmetric positive definite solution to the algebraic Riccati equation (ARE) is asymptotically stable for the DMRE, backward in time. In section 3, we further improve this result by showing that the DMRE also admits a linear L^2 gain [56] for any arbitrarily large set of initial conditions within the region of attraction, which we will refer to as "semi-global gain assignment." This conclusion lays the foundation of our small-gain analysis on the continuous-time VI [14, 13], which in turn leads to a sequence of convergence and robustness results. To better illustrate the differences among those aforementioned DP methods, we summarize their key features in Table 1. We admit that one drawback of robust DP is that it requires the nominal value of the model parameters in the algorithm, and hence is not model-free as are existing RL and adaptive DP (ADP) methods. This drawback can be easily conquered by combining our robust DP with existing adaptive optimal control results.

To demonstrate the power of the proposed method, in section 4, we apply robust DP to solve two classical problems arising from the field of adaptive optimal control. In the first application, we develop a continuous-time stochastic ADP theory for a class of ergodic control problems, which generalizes the results of [12]. Different from the

Table 1
Comparison of different DP methods.

Application scenarios	DP	Adaptive DP	Robust DP
Ideal case	Yes	Yes	Yes
Static uncertainty	No	Yes	Yes
Dynamic uncertainty	No	No	Yes
Model free	No	Yes	No

stochastic approximation [33] and Monte Carlo methods [50, Chapter 5] in traditional RL, a new convergence analysis method based on the robust DP theory is proposed in the continuous-time setting to handle the complex nature of the continuous-time ergodic control problem. In our second application, we propose a novel decentralized VI algorithm for solving coupled AREs. The small-gain theory [26] is applied with our robust DP framework to provide a sufficient condition for the convergence analysis of coupled AREs. This result is especially useful in developing robust ADP algorithms [25, 23, 57, 58] and solving non-zero-sum differential games [47, 48].

To further illustrate the proposed result, we also give two practical simulation examples in section 5.

Notation. Throughout this paper, I denotes the identity matrix. $\mathbb{R}$ and $\mathbb{R}_+$ denote the set of real numbers and the set of nonnegative real numbers, respectively. $\mathbb{Z}_+$ denotes the set of nonnegative integers. $\mathcal{S}^n$ denotes the normed space of all n-by-n real symmetric matrices, equipped with the induced matrix norm. $\mathcal{S}^n_+ = \{P \in \mathcal{S}^n : P > 0\}$. $\oplus$ indicates the Kronecker sum. B_ε denotes an open ball centered at the origin with radius ε. For two vectors x, $y \in \mathbb{R}^n$, $\langle x, y\rangle_{L^2} = x^T y$. $|\cdot|$ denotes the Euclidean norm for vectors, or the induced matrix norm for matrices. For a matrix $M \in \mathbb{R}^{n\times m}$, $\text{vec}(M) = [M_1^T\ M_2^T\ \cdots\ M_m^T]^T$, where $M_i \in \mathbb{R}^n$ is the ith column of M. For any $M \in \mathcal{S}^n$, denote $\lambda_m(M)$ and $\lambda_M(M)$ as the minimum and maximum eigenvalues of M, respectively, and let $\text{vech}(M) = [M_{11}\ M_{12}\ \cdots\ M_{1n}\ M_{22}\ M_{23}\ \cdots\ M_{(n-1)n}\ M_{nn}]^T$, where $M_{ij} \in \mathbb{R}$ is the (i,j)th element of matrix M. Finally, $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product.

2. Preliminaries.

2.1. System description.

Consider the following linear time-invariant system:

$$\dot{x} = Ax + Bu, \tag{1}$$

where $x \in \mathbb{R}^n$ is the system state, $u \in \mathbb{R}^m$ is the control input, and $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are system matrices. Assume (A, B) is stabilizable.

Denote the cost corresponding to system (1) as

$$\mathcal{J}(x(0); u) = \int_0^\infty (x^T Q x + u^T R u) ds, \tag{2}$$

where $Q = Q^T \geq 0$, $R = R^T > 0$, and $(A, Q^{1/2})$ is observable. It is well known that $\mathcal{J}$ is minimized under the optimal controller $u^* = -K^* x$, where $K^* = R^{-1}B^T P^*$, with P^* the unique symmetric positive definite solution to the following ARE:

$$0 = A^T P^* + P^* A - P^* B R^{-1} B^T P^* + Q. \tag{3}$$

Moreover, $A - BK^*$ is Hurwitz.

Algorithm 1. Continuous-time value iteration.

```
Choose $P_0 = P_0^T > 0$. $k, q \leftarrow 0$.
loop
    $P_{k+1/2} \leftarrow P_k + h_k(A^T P_k + P_k A - P_k B R^{-1} B^T P_k + Q)$
    if $P_{k+1/2} > 0$ and $|P_{k+1/2} - P_k|/h_k < \bar{\varepsilon}$ then
        return $P_k$ as an approximation to $P^*$
    else if $|P_{k+1/2}| > q$ or $P_{k+1/2} \ngtr 0$ then
        $P_{k+1} \leftarrow P_0$. $q \leftarrow q + 1$.
    else
        $P_{k+1} \leftarrow P_{k+1/2}$
    $k \leftarrow k + 1$
```

2.2. DMRE and continuous-time VI.

Since (3) is a nonlinear matrix equation, it is not easy to solve P^* from the ARE directly. One way of finding P^* is to use the continuous-time VI [14]. Before introducing the VI algorithm, we define a real sequence $\{h_k\}_{k=0}^{\infty}$ satisfying

$$h_k > 0, \quad \lim_{k\to\infty} h_k = 0, \quad \sum_{k=0}^{\infty} h_k = \infty.$$

Let $\bar{\varepsilon} > 0$ be a small threshold.

The continuous-time VI is recalled from [14] and shown in Algorithm 1. Note that if $Q > 0$, then the initial choice on P_0 can be relaxed to $P_0 = P_0^T \geq 0$. Detailed convergence analysis on Algorithm 1, and its extensions to model-free adaptive optimal controller design, can be found in [14]. However, how robust Algorithm 1 is to various types of disturbances still remains an open problem. As shown in subsequent sections, we will provide the first solution to this fundamentally challenging issue for continuous-time dynamical systems.

3. Robust DP and VI for continuous-time systems.

The purpose of this section is to extend Algorithm 1 in different directions by providing a concrete stability and robustness analysis for the DMRE and VI.

3.1. Robust DP and DMRE.

As it has been shown in [59, 32], for any $P(0) = P(0)^T \geq 0$, the solution to the following DMRE converges to P^* asymptotically as t goes to infinity:

$$\dot{P} = A^T P + PA - PBR^{-1}B^T P + Q. \tag{4}$$

Denoting $K = R^{-1}B^T P$, we have from (4) that

$$\begin{aligned}
\dot{P} &= A^T P + PA - PBR^{-1}B^T P + Q \\
&= (A - BK)^T P + P(A - BK) + K^T R K + Q \\
&= (A - BK^*)^T P + P(A - BK^*) + (K^*)^T R K^* + Q - (K - K^*)^T R (K - K^*).
\end{aligned}$$

Subtracting (3) from the above equation, and letting $\tilde{P} = P - P^*$, we have

$$\dot{\tilde{P}} = (A - BK^*)^T \tilde{P} + \tilde{P}(A - BK^*) - \tilde{P}BR^{-1}B^T\tilde{P}. \tag{5}$$

The following two lemmas play an important role in developing our robust VI.

Lemma 3.1. *$\hat{P}$ is globally[1] exponentially stable at P^*, where $\hat{P}$ is the solution of the following system:*

$$\dot{\hat{P}} = (A - BK^*)^T \hat{P} + \hat{P}(A - BK^*) + (K^*)^T RK^* + Q, \quad \hat{P}(0) \in \mathbb{R}^{n\times n}. \tag{6}$$

Proof. Denote $\xi = \mathrm{vec}(\hat{P} - P^*) \in \mathbb{R}^{n^2}$. Then, by subtracting (3) from (6), one has

$$\dot{\xi} = ((A - BK^*) \oplus (A - BK^*))^T \xi. \tag{7}$$

Since $A - BK^*$ is Hurwitz, $(A - BK^*) \oplus (A - BK^*)$ is also Hurwitz [16]. This completes the proof. □

Remark 3.2. Note from (6) that when $\hat{P}(0) \in \mathcal{S}^n$, we have $\hat{P}(t) \in \mathcal{S}^n$ for all $t > 0$. Since $\mathcal{S}^n \subset \mathbb{R}^{n\times n}$, we know $\hat{P}$ is also exponentially stable at P^* in $\mathcal{S}^n$.

The following lemma is a direct extension of the converse Lyapunov theorem.

Lemma 3.3. *Consider a dynamical system defined on $\mathcal{S}^n$:*

$$\dot{P} = G(P), \tag{8}$$

where $G : \mathcal{S}^n \to \mathcal{S}^n$ is locally Lipschitz, and satisfies $G(0) = 0$. If system (8) is asymptotically stable at the origin, and its region of attraction is R_A, then there exists a smooth Lyapunov function $V : R_A \to \mathbb{R}_+$, such that

$$\begin{aligned} &\langle \partial_x V(P), G(P)\rangle_F < 0, \quad V(P) > 0 \quad \forall P \in R_A \setminus \{0\}, \\ &\lim_{P\to\partial R_A} V(P) = \infty, \quad \langle \partial_x V(0), G(0)\rangle_F = 0, \quad V(0) = 0. \end{aligned}$$

Proof. Denote a mapping $\mathcal{M}(\cdot) : (\mathcal{S}^n, \langle\cdot,\cdot\rangle_F) \to (\mathbb{R}^{n(n+1)/2}, \langle\cdot,\cdot\rangle_{L^2})$, such that

$$\mathcal{M}(M) = [M_{11}\ \sqrt{2}M_{12}\ \cdots\ \sqrt{2}M_{1n}\ M_{22}\ \sqrt{2}M_{23}\ \cdots\ \sqrt{2}M_{(n-1)n}\ M_{nn}]^T.$$

Then, for any $M_1, M_2 \in \mathcal{S}^n$, $\mathcal{M}^T(M_1)\mathcal{M}(M_2) = \langle M_1, M_2\rangle_F$. Hence, $\mathcal{M}(\cdot)$ is a smooth isometric isomorphism.[2] Then, one can rewrite (8) as the following ODE:

$$\dot{p} = g(p), \tag{9}$$

where $p = \mathcal{M}(P)$, and $g = \mathcal{M} \circ G \circ \mathcal{M}^{-1}$. For system (9), denote the region of attraction of $0 \in \mathbb{R}^{n(n+1)/2}$ by R'_A. By the converse Lyapunov theorem [28, Theorem 4.17], we know there exists a smooth function $W(\cdot) : R'_A \to \mathbb{R}_+$, such that

$$\begin{aligned} &\partial_x W(p)g(p) < 0, \quad W(p) > 0 \quad \forall p \in R'_A \setminus \{0\}, \\ &\lim_{p\to\partial R'_A} W(p) = \infty, \quad \partial_x W(0)g(0) = 0, \quad W(0) = 0. \end{aligned}$$

We claim $R'_A = \mathcal{M}(R_A)$. Otherwise, if there exist $P_0 \in R_A$ and $\mathcal{M}(P_0) \notin R'_A$, then R'_A is no longer the region of attraction for (9) since the solution to (9) starting from $\mathcal{M}(P_0)$ also converges to the origin, by the norm preserving property of $\mathcal{M}$.

[1] This is global in the sense that the region of attraction is the entire normed space of all n-by-n real matrices equipped with the induced matrix norm.

[2] A bounded linear operator is called an isometric isomorphism if it is a norm preserving bijection which is continuous and has a continuous inverse [42, p. 71].

Similarly, if there exists $p_0 \in R'_A$ such that $\mathcal{M}^{-1}(p_0) \notin R_A$, then R_A is no longer the region of attraction for (8).

Now, we define a function $V(\cdot) : R_A \to \mathbb{R}_+$, such that $V = W \circ \mathcal{M}$. By the definition of matrix calculus, $(\partial_x V)_{i,j} \propto (\mathcal{M}^{-1} \circ \partial_x^T W \circ \mathcal{M})_{i,j}$. It is easy to see that all the higher-order derivatives of V can be defined in a similar manner. Hence, V is also smooth. By the definition of Frobenius inner product,

$$\partial_x W(\mathcal{M}(P)) g(\mathcal{M}(P)) = \langle \partial_x V(P), G(P) \rangle_F \quad \forall P \in \mathcal{S}^n.$$

This concludes the proof. ∎

Remark 3.4. Lemma 3.3 extends the converse Lyapunov theorem for general nonlinear systems [28, Theorem 4.17] to $\mathcal{S}^n$, by exploring the equivalence between $(\mathcal{S}^n, \langle\cdot,\cdot\rangle_F)$ and $(\mathbb{R}^{n(n+1)/2}, \langle\cdot,\cdot\rangle_{L^2})$ resulting from the isometric isomorphism mapping defined in the proof. In addition, the converse statement of Lemma 3.3, i.e., the Lyapunov theorem for the stability of systems over $(\mathcal{S}^n, \langle\cdot,\cdot\rangle_F)$, can also be derived.

Following analysis similar to that of Lemma 3.3, one can also extend [28, Theorem 4.14] to $\mathcal{S}^n$. We omit the proof of this direct extension to avoid duplication.

LEMMA 3.5. *Suppose system* (8) *is exponentially stable at the origin. If G is continuously differentiable and $\partial_x G$ is bounded on $\mathcal{S}^n$, then there exist a smooth Lyapunov function $V : R_A \to \mathbb{R}_+$ and $C_i > 0$, $i = 1, 2, 3, 4$, such that*[3]

$$C_1|P|^2 \le V(P) \le C_2|P|^2, \quad \dot{V}(P) \le -C_3|P|^2, \quad |\partial_x V(P)| < C_4|P|.$$

PROPOSITION 3.6. *P is exponentially stable at P^* over $\mathcal{S}^n$.*

Proof. Note from Lemma 3.1 and Remark 3.2 that $\hat{P}$ is globally exponentially stable at P^*. Then, by Lemma 3.5, we can find a smooth Lyapunov function $V : \mathcal{S}^n \to \mathbb{R}_+$ and $C_i > 0$, $i = 1, 2, 3, 4$, such that

$$\begin{gathered} C_1|\hat{P} - P^*|^2 \le V(\hat{P} - P^*) \le C_2|\hat{P} - P^*|^2, \\ \dot{V}(\hat{P} - P^*) \le -C_3|\hat{P} - P^*|^2, \quad |\partial_x V(\hat{P} - P^*)| < C_4|\hat{P} - P^*| \quad \forall \hat{P} \in \mathcal{S}^n. \end{gathered}$$

Comparing the dynamics of $\hat{P}$ and P, we see that the only difference between these two systems is the quadratic term $\tilde{P}BR^{-1}B^T\tilde{P}$. Now, by taking the derivative of V along the solutions of system (5), we have

$$\dot{V}(\tilde{P}) \le -C_3|\tilde{P}|^2 + C_5|\partial_x V(\tilde{P})||\tilde{P}|^2 \le -C_3|\tilde{P}|^2 + C_4C_5|\tilde{P}|^3 \quad \forall \tilde{P} \in \mathcal{S}^n,$$

where $C_5 > 0$ is a scalar that depends on R and B. From the above inequality, we know there exists $C_6 > 0$, such that

$$\dot{V}(\tilde{P}) \le -C_6|\tilde{P}|^2 \le -\frac{C_6}{C_2}V(\tilde{P}) \quad \forall |\tilde{P}| \le \frac{C_3}{C_4C_5}.$$

The proof is then completed using the Lyapunov theorem [28, Theorem 4.10]. ∎

We will exploit the important feature of exponential stability further in the rest of this paper. First, let us consider the following variant of (4) subject to a disturbance input $\Delta(t) = \Delta^T(t)$:

$$\dot{P}_\Delta = A^T P_\Delta + P_\Delta A - P_\Delta BR^{-1}B^T P_\Delta + Q + \Delta, \quad P_\Delta(0) = P_\Delta^T(0) \ge 0. \tag{10}$$

[3] Note that we use the induced norm here instead of the Frobenius norm in Lemma 3.3, due to the equivalence of matrix norms.

Remark 3.7. Δ can represent a large class of disturbances. In particular, we conduct robustness analysis associated with (10) in Theorem 3.8 below by considering three different forms of Δ, including (a) a bounded state-independent external signal (Theorem 3.8, parts (i) and (ii)); (b) the output of an input-to-output stable (IOS) [46, section 7] nonlinear dynamical system (Theorem 3.8, part (iii)); and (c) a stochastic noise (Theorem 3.8, part (iv)). The assumption $\Delta(t) = \Delta^T(t)$ is to guarantee that P_Δ is always symmetric. This condition can be easily satisfied in practice, since for any $M \in \mathbb{R}^{n\times n}$, $x^T M x = \frac{1}{2}x^T(M + M^T)x$, and $\frac{1}{2}(M + M^T)$ is real symmetric.

THEOREM 3.8. *Consider system* (10) *with* $Q > 0$. *Denoting* $\tilde{P}_\Delta = P_\Delta - P^*$, *we have the following:*

(i) *If* $\inf_t \lambda_m(Q + \Delta(t)) \geq 0$ *and* $\sup_t \lambda_M(Q + \Delta(t)) < \infty$, *then* P_Δ *is well defined on* $\mathbb{R}_+$, *and there exists* $M \in \mathcal{S}^n$ *that is dependent on* $P_\Delta(0)$, *such that* $0 \leq P_\Delta(t) < M$ *for all* $t > 0$.

(ii) *If* Δ *satisfies the conditions in* (i), *and* $\lim_{t\to\infty} \Delta(t) = 0$, *then* $\lim_{t\to\infty} P_\Delta(t) = P^*$. *If, in addition,* $\Delta \in L^2$, *then* $\tilde{P}_\Delta \in L^2$.

(iii) *There exists* $\gamma > 0$, *such that if the system*[4]

$$\dot{M} = f(M, P_\Delta), \quad \Delta(t) = \Delta(P_\Delta, M), \tag{11}$$

where f *and* Δ *are locally Lipschitz,* $f(M^*, P^*) = 0$, *and* $\Delta(P^*, M^*) = 0$, *is zero-state detectable*[5] *and admits an IOS Lyapunov function* V_f *satisfying*

$$\dot{V}_f(\tilde{M}) \leq -|\Delta|^2 + \gamma^2|\tilde{P}_\Delta|^2 \quad \forall M \in B_{\varepsilon_0}(M^*), \tag{12}$$

for some $\varepsilon_0 > 0$, *where* $\tilde{M} = M - M^*$, *then* (P_Δ, M) *is asymptotically stable at* (P^*, M^*).

(iv) *Suppose* $\Delta(t) = \sum_{i=1}^N \Delta_i(P_\Delta)v_i(t)$, *where* $N > 0$, $\Delta_i : \mathcal{S}^n \to \mathcal{S}^n$, *and the* v_i *are one-dimensional independent and identically distributed Gaussian white noises. Then, there exists* $\gamma > 0$, *such that if* $\sum_i |\Delta_i|^2 < \gamma|\tilde{P}_\Delta|^2$ *in a neighborhood of* P^*, *then* P_Δ *is asymptotically stable at* P^* *in the mean square sense.*

The proof of Theorem 3.8 is given in Appendix A.1.

Proposition 3.6 and Theorem 3.8 imply that the DMRE behaves very similarly to an exponentially stable linear system in a neighborhood of P^*, and thus exhibits a series of nice properties. However, the stability and robustness results in this proposition and this theorem are of limited use in practice, since they hold only in a neighborhood of P^*. In order to obtain desirable transient performance for the DMRE in a sufficiently large compact set, we need to design carefully the cost (2). Indeed, the following corollary shows that by choosing Q and R properly, we can guarantee the *semi-global* exponential stability of (4) at P^*. By "semi-global," we mean that the domain of attraction is bounded but can be made as large as possible [43].

COROLLARY 3.9. *Given* $Q_0 = Q_0^T > 0$ *and* $R_0 = R_0^T > 0$, *for any compact set* $\mathcal{S}_0 \subset \mathcal{S}_+^n$, *there exists a constant* $\lambda > 0$, *such that by choosing* $Q = \lambda Q_0$ *and* $R = \lambda R_0$, *each trajectory of* (4) *starting at* $P(0) \in \mathcal{S}_0$ *converges exponentially to* P^*.

[4] M can be either a real vector or a real matrix, depending on the specific problem formulation. For consistency, here we consider M as a real matrix of an appropriate dimension.

[5] Here, with slight abuse of notation, we say (11) is zero-state detectable if $\Delta \equiv 0$ and $P_\Delta \equiv P^*$ imply $M \equiv M^*$.

Proof. First, note that under the choice of $Q = \lambda Q_0$ and $R = \lambda R_0$, K^* is independent of λ, as both Q and R are derived from Q_0 and R_0 by multiplying the same scaling factor. Moreover, P^* is a linear function of λ, and $\lim_{\lambda \to 0^+} P^* = 0$. Now, for any $\mathcal{S}_0$, we can find a small enough $\lambda > 0$, such that $P^* < P(0)$ for all $P(0) \in \mathcal{S}_0$. Then, by choosing $\hat{P}(0) = P(0)$ in (6), we have for any given $t > 0$ and $x(-t) \in \mathbb{R}^n$,

$$x^T(-t)P(t)x(-t) = \inf_u \left\{ x^T(0)P(0)x(0) + \int_{-t}^{0} (x^TQx + u^TRu)ds \right\}$$
$$\leq (x^*(0))^T P(0)x^*(0) + \int_{-t}^{0} (x^*)^T(Q + (K^*)^T RK^*)x^* ds = x^T(-t)\hat{P}(t)x(-t),$$

where x^* is the solution to system (1) with $u = -K^*x^*$ and $x^*(-t) = x(-t)$. Moreover, by monotonicity [13, Lemma 1], $P^* \leq P(t)$ for all t. Since by Lemma 3.1 $\hat{P}(t)$ converges to P^* exponentially, $x^T\hat{P}x$ also converges to x^TP^*x exponentially for all x. Using $x^TP^*x \leq x^TPx \leq x^T\hat{P}x$, we know x^TPx converges to x^TP^*x exponentially. Noting that this is true for all x, P thus converges to P^* exponentially. This completes the proof. □

Remark 3.10. It is easy to see from Corollary 3.9 that although multiplying the same scalar to Q_0 and R_0 does not influence the optimal feedback gain matrix, the transient performance of the DMRE can be quite different. Given any $P(0)$, by Proposition 3.6 and the converse Lyapunov theorem [28, Theorem 4.14], we can find a Lyapunov function V satisfying

$$C_1|\tilde{P}|^2 \leq V(\tilde{P}) \leq C_2|\tilde{P}|^2, \quad \dot{V}(\tilde{P}) \leq -C_3|\tilde{P}|^2, \quad |\partial_x V(\tilde{P})| < C_4|\tilde{P}|,$$

where $C_i > 0$, $i = 1, 2, 3, 4$, over a connected compact set including $P(0)$ and P^*. As a result, Corollary 3.9 allows us to extend the result obtained in Theorem 3.8 to any compact sets containing P^* in $\mathcal{S}^n_+$.

If we are allowed to have more freedom in choosing Q and R, it is possible to have the following semi-global gain assignment result.

COROLLARY 3.11. *Given $Q_0 = Q_0^T > 0$ and $R_0 = R_0^T > 0$, if B has full rank, then for any $\varepsilon > 0$ and $\gamma > 0$, there exists a sufficiently large $\lambda > 0$, such that for Δ satisfying* $\inf_t \lambda_m(Q + \Delta(t)) \geq 0$, (10) *admits a finite linear L^2 gain from Δ to $\tilde{P}_\Delta$ less than or equal to γ for $P(0) \in \{P \in \mathcal{S}^n : P \in B_\varepsilon(P^*)\}$, with $Q = \lambda Q_0$ and $R = R_0$.*

Proof. Since only Q_0 is multiplied by the scaling factor λ, different from Corollary 3.9, the optimal controller depends on λ here. Hence, the first step of our proof is to characterize the influence of λ on the optimal controller.

Note that choosing $Q = \lambda Q_0$ and $R = R_0$ is equivalent to choosing $Q = Q_0$ and $R = \lambda^{-1}R_0$, in the sense that these two choices lead to the same optimal controller. Denoting P^*_λ as the solution to (3) with $Q = Q_0$ and $R = \lambda^{-1}R_0$, we have

$$(A - BK^*_\lambda)^T P^*_\lambda + P^*_\lambda(A - BK^*_\lambda) = -\left(Q_0 + \left(\sqrt{\lambda}P^*_\lambda\right) BR_0^{-1}B^T \left(\sqrt{\lambda}P^*_\lambda\right)\right), \tag{13}$$

where $K^*_\lambda = \lambda R_0^{-1}B^TP^*_\lambda$. Since B has full rank, we know from [34, eq. (40)] that there exists $\bar{P} \in \mathcal{S}^n$, such that $\lim_{\lambda \to \infty} \sqrt{\lambda}P^*_\lambda = \bar{P}$. Thus, for any two positive constants C and ε, we can choose a large enough λ, such that $\left|\sqrt{\lambda}P^*_\lambda - \bar{P}\right| < \varepsilon$ and

$$\sqrt{\lambda}\left(Q_0 + \left(\sqrt{\lambda}P^*_\lambda\right) BR_0^{-1}B^T \left(\sqrt{\lambda}P^*_\lambda\right)\right) > CI.$$

This, together with the Lyapunov equation (13), implies that for any $\alpha > 0$, we can find $\lambda > 0$, such that

$$(A - BK_\lambda^*)^T M + M(A - BK_\lambda^*) < -\alpha M$$

for some constant matrix $M = M^T > 0$. This suggests that the eigenvalues of $A - BK_\lambda^*$ (and hence $(A - BK_\lambda^*) \oplus (A - BK_\lambda^*)$) can be placed arbitrarily far to the left from the imaginary axis in the complex plane, by choosing a large enough λ.

Now we come back to our original problem with $Q = \lambda Q_0$ and $R = R_0$. By robust pole placement [19], we know that for any $\gamma > 0$, one can find a $\lambda > 0$, such that the following system admits a linear L^2 gain from $\bar{\Delta}$ to $\tilde{P}_\Delta$ less than or equal to γ:

$$\dot{\tilde{P}}_\Delta = (A - BK_\lambda^*)^T \tilde{P}_\Delta + (A - BK_\lambda^*)\tilde{P}_\Delta + \bar{\Delta}(\tilde{P}_\Delta),$$

where $\tilde{P}_\Delta = P_\Delta - P^*$ and $\bar{\Delta}(\tilde{P}_\Delta) = \Delta - \tilde{P}_\Delta BR^{-1}B^T\tilde{P}_\Delta$. Comparing the above system with (5), and following similar arguments in the proof of Proposition 3.6, we know that for any $\varepsilon > 0$, the L^2 gain of system (10) can be made arbitrarily small on $\{P \in \mathcal{S}^n : P \in B_\varepsilon(P^*)\}$, by choosing a sufficiently large λ. This completes the proof. □

Remark 3.12. The full-rank condition on B is required to satisfy the matching condition, which is a common assumption in nonlinear gain assignment and robust control literature [27, 41, 21, 38]. To relax this assumption in the case of unmatched disturbance, one way is to study cascaded systems with full-rank input matrices via recursive backstepping [31], or a combination of the backstepping and small-gain approaches [38].

The following corollary is a direct extension of Theorem 3.8, parts (iii) and (iv), and Corollary 3.11, and thus its proof is omitted.

Corollary 3.13. *Given $Q_0 = Q_0^T > 0$, $R_0 = R_0^T > 0$, and $\lambda > 0$, define $Q = \lambda Q_0$ and $R = R_0$. Suppose B has full rank.*

(i) *For any $\gamma > 0$, if system* (11) *satisfies the conditions in Theorem* 3.8*, part* (iii)*, then there exists $\lambda > 0$, such that (P_Δ, M) is asymptotically stable at (P^*, M^*).*

(ii) *For any $\gamma > 0$ and $\varepsilon > 0$, if Δ satisfies the definition in Theorem* 3.8*, part* (iv)*, then there exists $\lambda > 0$, such that P_Δ is asymptotically stable at P^* in the mean square sense.*

3.2. Robust VI algorithm.

In this subsection, we formally introduce the robust VI algorithm (Algorithm 2) based on the theoretical results developed in subsection 3.1. Note that unlike Algorithm 1, Algorithm 2 includes both a deterministic perturbation term Δ_k and a stochastic noise term W_k in the updating equation of P_k.

The following theorem shows that Algorithm 2 inherits the robustness property from (10).

Theorem 3.14. *Denote a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $\{\mathcal{F}_k\}_{k\in\mathbb{Z}_+}$. Suppose $Q > 0$, W_k is $\mathcal{F}_k$-adapted, h_k is a sequence satisfying the conditions in subsection* 2.2*, and $\sum_{k=0}^\infty h_k W_k$ converges with probability one. Given $\{P_k\}_{k=0}^\infty$ defined in Algorithm* 2*, we have with probability one that*

(i) *there exist $\delta_0 > 0$, $N \geq 0$, and a compact set $\mathcal{S}_0 \subset \mathcal{S}_+^n$ with nonempty interior and $P^* \in \mathcal{S}_0$, such that if $|\Delta_k| < \delta_0(1 + |P_k|)$, then $\{P_k\}_{k=N}^\infty \subset \mathcal{S}_0$;*

(ii) *if $\lim_{k\to\infty} \Delta_k = 0$, then $\lim_{k\to\infty} P_k = P^*$;*

Algorithm 2. Continuous-time robust value iteration.

```
Choose $P_0 = P_0^T \geq 0$. $k, q \leftarrow 0$.
loop
    $P_{k+1/2} \leftarrow P_k + h_k(A^T P_k + P_k A - P_k B R^{-1} B^T P_k + Q + \Delta_k + W_k)$
    if $|P_{k+1/2}| > q$ or $P_{k+1/2} \not> 0$ then
        $P_{k+1} \leftarrow P_0$. $q \leftarrow q + 1$.
    else
        $P_{k+1} \leftarrow P_{k+1/2}$
    $k \leftarrow k + 1$
```

(iii) *if $\Delta_k := \Delta(P_k, M_k)$ is the output to the updating equation*

$$M_{k+1} = M_k + h_k f(M_k, P_k) + Z_k, \tag{14}$$

where $\{M_k\}_{k=0}^{\infty}$ is bounded in $B_{\varepsilon_0}(M^*)$ under a projection term Z_k for some $\varepsilon_0 > 0$, then there exists $\gamma > 0$, such that if the conditions in part (iii) *of Theorem* 3.8 *are satisfied, we have* $\lim_{k\to\infty}(P_k, M_k) = (P^*, M^*)$ *locally.*

The proof of Theorem 3.14 is given in Appendix A.2.

Remark 3.15. The first two parts of Theorem 3.14 focus on handling static uncertainties represented by either a bounded external disturbance input or a bounded function of P_k. Part (iii) of Theorem 3.14 deals with dynamic uncertainty and hence is more suitable for developing decentralized VI algorithms (see subsection 4.2).

Remark 3.16. The boundedness of M_k can be relaxed by extending the projection term in (14) to the adaptive boundary case as in Algorithm 2 [18, 33]. In addition, the local convergence result in part (iii) of Theorem 3.14 can be strengthened to the semi-global case by using the gain assignment method developed in Corollary 3.11.

The following corollary plays an important role in developing adaptive optimal control methods on the basis of the proposed robust VI framework. The proof of Corollary 3.17 is given in Appendix A.3.

Corollary 3.17. *Denote a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $\{\mathcal{F}_k\}_{k\in\mathbb{Z}_+}$. Consider Algorithm* 2 *with $W_k = \sigma(P_k)v_k$, $\Delta_k = \Delta_k(P_k)$, and $\sum_{k=0}^{\infty} h_k^2 < \infty$, where $\lim_{k\to\infty}\Delta_k = 0$ uniformly on any compact set, σ_i are continuous, and v_k is an $\mathcal{F}_k$-adapted martingale difference with finite variance. Then,* $\lim_{k\to\infty} P_k = P^*$ *with probability one.*

4. Applications to adaptive/stochastic/decentralized optimal control.

In this section, we provide two applications of the above robust VI method in solving adaptive optimal control problems that appear intractable using traditional DP methods.

4.1. Stochastic ADP for ergodic control problems.

In this subsection, we develop an ADP algorithm to solve the ergodic control problem [15] for linear stochastic systems with additive noise.

Consider the following system:

$$dx = (Ax + Bu)\,dt + \sum_{i=1}^{q_x} \sigma_i dw_{x,i}, \tag{15}$$

$$udt = -K_0 xdt + \sum_{i=1}^{q_u} \sigma_{u,i} dw_{u,i}, \tag{16}$$

where x, u, A, and B follow the same definitions as in system (1); $x(0)$ is deterministic; $w_{x,i}$ and $w_{u,i}$ are independent Brownian motions; $q_x, q_u \in \mathbb{Z}_+$; K_0 is a known initial input matrix; and $\sigma_i \in \mathbb{R}^n$ and $\sigma_{u,i} \in \mathbb{R}^m$ are constant vectors satisfying $\sum_{i=1}^{q_x} \sigma_i \sigma_i^T > 0$ and $\sum_{i=1}^{q_u} \sigma_{u,i}\sigma_{u,i}^T > 0$.

Remark 4.1. $\sum_{i=1}^{q_x} \sigma_i dw_{x,i}$ in (15) represents the additive noise in system (15). $\sum_{i=1}^{q_u} \sigma_{u,i} dw_{u,i}$ in (16) serves as an exploration noise, which has been widely used in adaptive control literature to guarantee the persistent excitation (PE) condition [52, Definition 3.2]. Note that besides the Brownian motion, other types of exploration noises can also be used. For simplicity, we only consider inputs in the form of (16) here, as in this case system (15) is purely driven by Brownian motions, and several standard results from stochastic analysis theory can be applied directly.

Assumption 4.2. *There exists an ergodic stationary probability measure μ on $\mathbb{R}^n \times \mathbb{R}^m$ for system* (15)–(16).

An analogue of Assumption 4.2 for MDPs has been widely used in the approximate DP and RL literatures [7, 50, 54, 55]. For conditions on the existence and uniqueness of the stationary distribution, see [29, Chapter 4].

The objective of ergodic control is to minimize (with probability one)

$$\mathcal{J}(u) = \limsup_{T\to\infty} \frac{1}{T} \int_0^T (x^T Q x + u^T R u) dt,$$

where $Q = Q^T > 0$ and $R = R^T > 0$. It can be shown [15] that $\inf_u \mathcal{J}(u) = \sum_{i=1}^{q_3} \sigma_i^T P^* \sigma_i$, with P^* and the optimal controller sharing the same definitions as the ones in subsection 2.1 for deterministic systems.

Now, we derive an online ADP algorithm to solve the above ergodic control problem. For all $x \in \mathbb{R}^n$ and $P \in \mathcal{S}^n$, by Itô's lemma [49, Theorem 8.3], we have along the trajectories of (15) that

$$\begin{aligned} d(x^T P x) &= 2x^T P(Ax + Bu)dt + \sum_{i=1}^{q_x} \sigma_i^T P \sigma_i dt + 2x^T P \sum_{i=1}^{q_x} \sigma_i dw_{x,i} \\ &= \psi^T(z)\theta(P)dt - r(z)dt + 2x^T P \sum_{i=1}^{q_x} \sigma_i dw_{x,i}, \end{aligned} \tag{17}$$

where $z = [x^T \;\; u^T \;\; 1]^T$, $r(z) = x^T Q x + u^T R u$,

$$\begin{aligned} \psi(z) &= [z_1^2 \; 2z_1 z_2 \; \cdots \; 2z_1 z_{n+m+1} \; z_2^2 \; 2z_2 z_3 \; \cdots \; 2z_{n+m} z_{n+m+1} \; z_{n+m+1}^2]^T, \\ \theta(P) &= \operatorname{vech}\left(\begin{bmatrix} PA + A^T P + Q & PB & 0 \\ B^T P & R & 0 \\ 0 & 0 & \sum_{i=1}^{q_x} \sigma_i^T P \sigma_i \end{bmatrix} \right). \end{aligned}$$

Then, multiplying by ψ on both sides of (17), we have on any finite time interval $[0,T]$ that

$$\begin{aligned}\frac{1}{T}\int_0^T \psi\psi^T dt\theta(P) = \frac{1}{T}\int_0^T \psi d(x^TPx) + \frac{1}{T}\int_0^T \psi r dt \\ -\frac{2}{T}\int_0^T \psi x^T P\sum_{i=1}^{q_x}\sigma_i dw_{x,i}.\end{aligned} \tag{18}$$

Once $\theta(P)$ is solved from the equation above, we can define a transformation $\mathcal{T}$, such that $\mathcal{R}(P) = \mathcal{T}(\theta(P))$, where $\mathcal{R}$ is defined in Appendix A.2.

In order to solve (18), we impose the following assumption.

Assumption 4.3. *There exist $t_0, c > 0$, such that for all $t \geq t_0$,*

$$\frac{1}{t}\int_0^t \psi\psi^T ds \geq cI \quad \textit{with probability one.} \tag{19}$$

Note that system (15)–(16) is a multidimensional Ornstein–Uhlenbeck process. Hence, its stationary probability measure μ is also Gaussian, and thus (x,u) has finite rth moment for any $r \in \mathbb{Z}_+$. Assumption 4.3 is similar to the PE condition widely used in the adaptive control literature (see Remark 4.1 for details).

By a direct extension of Birkhoff's ergodic theorem [2, Theorem 1.5.18] and the Itô isometry [49, Theorem 6.1], we know[6]

$$\lim_{t\to\infty}\mathbb{E}^{\mathbb{P}}\left[\left\|\frac{1}{t}\int_0^t \psi\psi^T ds - \int_{\mathbb{R}^n\times\mathbb{R}^m}\psi\psi^T d\mu\right\|_2^2\right] = 0, \tag{20}$$

$$\begin{aligned}&\lim_{t\to\infty}\mathbb{E}^{\mathbb{P}}\left[\left\|\frac{1}{t}\int_0^t \psi d(x^TPx) + \frac{1}{t}\int_0^t \psi r ds - \frac{1}{t}\int_0^t \psi\psi^T ds\theta(P)\right\|_2^2\right] \\ &\qquad = \lim_{t\to\infty}\frac{4}{t^2}\mathbb{E}^{\mathbb{P}}\left[\left\|\int_0^t \psi x^T P\sum_{i=1}^{q_x}\sigma_i dw_{x,i}\right\|_2^2\right] = 0.\end{aligned} \tag{21}$$

Choosing a monotonically increasing sequence $\{t_k\}_{k=0}^\infty$ with t_0 satisfying conditions in Assumption 4.3 and $\lim_{k\to\infty} t_k = \infty$, we denote

$$\hat{\theta}(P,t_k) = \left(\int_0^{t_k}\psi\psi^T ds\right)^{-1}\left(\int_0^{t_k}\psi d(x^TPx) + \int_0^{t_k}\psi r ds\right).$$

For simplicity, denote $\hat{\theta}_k = \hat{\theta}(P_k, t_k)$. The VI-based ADP algorithm for the ergodic control problem is given in Algorithm 3.

Theorem 4.4. *Under Assumptions* 4.2 *and* 4.3*, we have $\lim_{k\to\infty} P_k = P^*$ with probability one, where $\{P_k\}_{k=0}^\infty$ is obtained from Algorithm* 3.

Proof. First, for any $P \in \mathcal{S}^n$, denote

$$\Delta_k(P) := \hat{\theta}(P,t_k) - \theta(P) = 2\left(\int_0^{t_k}\psi\psi^T ds\right)^{-1}\int_0^{t_k}\psi x^T P\sum_{i=1}^{q_x}\sigma_i dw_{x,i}.$$

[6] $\|\cdot\|_2$ denotes the matrix 2-norm. $\mathbb{E}^{\mathbb{P}}$ is the expectation on a probability space $(\Omega,\mathcal{F},\mathbb{P})$, where Ω is a sample space, $\mathcal{F}$ is a σ-field of Borel sets of Ω, and $\mathbb{P}$ is a stationary distribution of (x,u) such that $\int_\Omega f(x(\omega),u(\omega))d\mathbb{P}(\omega) = \int_{\mathbb{R}^n\times\mathbb{R}^m} f(x,u)d\mu(x,u)$ for all measurable f.

Algorithm 3. Online robust optimal control design for ergodic control.

```
Choose $P_0 = P_0^T \geq 0$. $k, q \leftarrow 0$. Pick an input $u$ in form of (16).
loop
    $\hat{\theta}_k \leftarrow \big(\int_0^{T_k} \psi\psi^T dt\big)^{-1}\big(\int_0^{T_k} \psi d(x^T P_k x) + \int_0^{T_k} \psi r dt\big)$.
    $P_{k+1/2} \leftarrow P_k + h_k \mathcal{T}(\hat{\theta}_k)$
    if $P_{k+1/2} > 0$ and $|P_{k+1/2} - P_k|/h_k < \bar{\varepsilon}$ then
        return $P_k$ as an approximation to $P^*$
    else if $|P_{k+1/2}| > q$ or $P_{k+1/2} \not> 0$ then
        $P_{k+1} \leftarrow P_0$. $q \leftarrow q + 1$.
    else
        $P_{k+1} \leftarrow P_{k+1/2}$
    $k \leftarrow k + 1$
```

Then, by (20) and (21), we have $\lim_{k\to\infty} \mathbb{E}^{\mathbb{P}}\left[\|\Delta_k(P)\|_2^2\right] = 0$. Hence, $\Delta_k(P)$ is a martingale (elementwise). By Burkholder–Davis–Gundy inequality [17, Theorem 1.1], we have

$$\mathbb{E}^{\mathbb{P}}\left[|\Delta_k^{i,j}(P)|^4\right] \leq C\mathbb{E}^{\mathbb{P}}\left[[\Delta^{i,j}(P)]_k^2\right]$$

for some constant $C > 0$, where $\Delta_k^{i,j}$ is the (i,j)th element of Δ_k, and $[\cdot]_k$ denotes the quadratic variation [49, section 8.6]. By the fact that (x,u) has finite rth moment for any $r \in \mathbb{Z}_+$, we have $\sup_{k\geq 0} \mathbb{E}^{\mathbb{P}}\left[[\Delta^{i,j}(P)]_k^2\right] < \infty$. This implies that the variance of $\Delta_k^T\Delta_k$ is bounded.

Now, the updating equation in Algorithm 3 is equivalent to

$$P_{k+1/2} \leftarrow P_k + h_k(\mathcal{R}(P_k) + \Delta_{1,k}(P_k) + \Delta_{2,k}(P_k)),$$

where $\Delta_{1,k}(P_k)$ is a zero-mean stochastic noise with finite variance for each k, and $\Delta_{2,k}(\cdot)$ is deterministic and decreases to 0 as k goes to infinity. The proof is then completed by Corollary 3.17. ∎

Remark 4.5. The ADP method developed in this subsection can be extended to solve the discount optimal control problem [12], by exploring the relationship between the discounted cost and the ergodic cost [12, Lemma 4].

4.2. Decentralized VI. In previous sections, we have studied different types of optimal control problems for continuous-time linear systems. A common feature of these results is that the optimal controller and value function can be obtained by solving a single ARE. However, in some applications, including the non-zero-sum differential game and the robust ADP, the optimal solution is solved from a group of cascaded or coupled AREs/HJB equations. Here, we present a decentralized VI framework for continuous-time linear systems based on the robust VI proposed in section 3.

For simplicity, let us consider a network of two agents, with each agent i, $i = 1, 2$, aiming at solving a linear optimal control problem (see subsection 2.1) defined by four matrices (A_i, B_i, Q_i, R_i). Obviously, if (A_1, B_1, Q_1, R_1) and (A_2, B_2, Q_2, R_2) are not dependent on each other, then each agent can solve its own optimal control problem independently. However, assuming now that agent i's system information (A_i, B_i, Q_i, R_i) depends on agent j's ($j \neq i$) optimal solution (P_j^*, K_j^*) through a nonlinear function $\Delta_i(\cdot)$, and that for security reasons the two agents cannot exchange their system information (A_i, B_i, Q_i, R_i), $i = 1, 2$, then it is no longer a trivial task how to solve (P_i^*, K_i^*) in a decentralized manner. Reformulating this problem mathematically, we focus on solving the following two coupled AREs:

$$0 = A_1^T P_1^* + P_1^* A_1 - P_1^* B_1 R_1^{-1} B_1^T P_1^* + Q_1 + \Delta_1(P_1^*, P_2^*),$$
$$0 = A_2^T P_2^* + P_2^* A_2 - P_2^* B_2 R_2^{-1} B_2^T P_2^* + Q_2 + \Delta_2(P_2^*, P_1^*),$$

where $(A_i, B_i, Q_i, R_i) \in \mathbb{R}^{n_i \times n_i} \times \mathbb{R}^{n_i \times m_i} \times \mathcal{S}_+^{n_i} \times \mathcal{S}_+^{m_i}$, and $\Delta_1 = \Delta_1^T$ and $\Delta_2 = \Delta_2^T$ are two continuous nonlinear functions.

ASSUMPTION 4.6. *There exist four polynomials $\gamma_{i,j} \in \mathcal{K}$, $i, j = 1, 2$, such that*[7]

$$|\tilde{\Delta}_1(P_1, P_2)| \leq \gamma_{1,1}(|\tilde{P}_1|) + \gamma_{1,2}(|\tilde{P}_2|), \quad |\tilde{\Delta}_2(P_2, P_1)| \leq \gamma_{2,2}(|\tilde{P}_2|) + \gamma_{2,1}(|\tilde{P}_1|),$$

where $\tilde{\Delta}_1(P_1, P_2) = \Delta_1(P_1, P_2) - \Delta_1(P_1^*, P_2^*)$, $\tilde{\Delta}_2(P_2, P_1) = \Delta_2(P_2, P_1) - \Delta_2(P_2^*, P_1^*)$, $\tilde{P}_1 = P_1 - P_1^*$, and $\tilde{P}_2 = P_2 - P_2^*$.

Remark 4.7. Assumption 4.6 holds widely in different control problems. For example, in two-player non-zero-sum differential games, we have $A_1 = A_2$ and

$$\Delta_i(P_i, P_j) = P_j B_j R_j^{-1} R_{ij} R_j^{-1} B_j^T P_j - P_j B_j R_j^{-1} B_j^T P_i - P_i B_j R_j^{-1} B_j^T P_j,$$

where $i \neq j$ and $R_{ij} = R_{ij}^T > 0$. Also, in the robust ADP design for systems with unmatched disturbances [24, Chapter 5.1.1.2], we have $\Delta_1 = 0$ and

$$\Delta_2(P_2, P_1) = P_2 R_1^{-1} B_1^T P_1 B_1 + B_1^T P_1 B_1 R_1^{-1} P_2.$$

Note that $\gamma_{i,j}$ may depend on P_1^* and P_2^*.

The following theorem provides a convergence analysis for the coupled DMREs using small-gain theory [26].

THEOREM 4.8. *Under Assumption* 4.6, *there exist $\varepsilon > 0$ and small enough $|\gamma_{i,j}|$, $i, j \in \{1, 2\}$, such that given $(P_1(0), P_2(0))$ in a ε-neighborhood of (P_1^*, P_2^*), we have $\lim_{t\to\infty} P_1(t) = P_1^*$ and $\lim_{t\to\infty} P_2(t) = P_2^*$, where*

$$\dot{P}_1 = A_1^T P_1 + P_1 A_1 - P_1 B_1 R_1^{-1} B_1^T P_1 + Q_1 + \Delta_1(P_1, P_2), \tag{22}$$
$$\dot{P}_2 = A_2^T P_2 + P_2 A_2 - P_2 B_2 R_2^{-1} B_2^T P_2 + Q_2 + \Delta_2(P_2, P_1). \tag{23}$$

Moreover, if B_1 and B_2 have full rank, the convergence result holds for any $\gamma_{i,j}$ and ε, by picking Q_1 and Q_2 properly.

Proof. Following the derivation of (26) in Appendix A.1, there exist $\varepsilon > 0$ and a Lyapunov function V, such that

$$\begin{aligned}\dot{V}(\tilde{P}_1, \tilde{P}_2) &\leq -C_1(|\tilde{P}_1|^2 + |\tilde{P}_2|^2) + C_2|\tilde{P}_1||\tilde{\Delta}_1| + C_3|\tilde{P}_2||\tilde{\Delta}_2| \\ &\leq -C_1(|\tilde{P}_1|^2 + |\tilde{P}_2|^2) + C_2|\tilde{P}_1| \sum_{j=1,2} \gamma_{1,j}(|\tilde{P}_j|) + C_3|\tilde{P}_2| \sum_{j=1,2} \gamma_{2,j}(|\tilde{P}_j|) \\ &\leq -\frac{C_1}{2}(|\tilde{P}_1|^2 + |\tilde{P}_2|^2) + C_4 \sum_{i,j=1,2} \gamma_{i,j}^2(|\tilde{P}_j|) \quad \forall |\tilde{P}_1| < \varepsilon,\ |\tilde{P}_2| < \varepsilon,\end{aligned}$$

where $C_i > 0$, $i = 1, 2, 3, 4$, are constants. Since $\gamma_{i,j}$ are polynomials, the second term on the right-hand side of the above inequality decreases to 0 at least as fast as the

[7] A function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ is of class $\mathcal{K}$ if it is continuous, strictly increasing, and $\gamma(0) = 0$.

Algorithm 4. Decentralized value iteration.

```
For the $i$th subsystem, choose $P_{i,0} = P_{i,0}^T \geq 0$. $k \leftarrow 0$.
loop
    $P_{i,k+1} \leftarrow P_{i,k} + h_{i,k}(A_i^T P_{i,k} + P_{i,k} A_i - P_{i,k} B_i R_i^{-1} B_i^T P_{i,k} + Q_i + \Delta_i(P_{i,k}, P_{j,k}))$
    if $|P_{i,k+1} - P_{i,k}|/h_{i,k} < \bar{\varepsilon}$ then
        return $P_{i,k}$ as an approximation to $P_i^*$
    $k \leftarrow k + 1$
```

first term. Hence, (22) and (23) are asymptotically stable at (P_1^*, P_2^*), as long as the gain of $\gamma_{i,j}$ is small enough.

Moreover, if B_1 and B_2 have full rank, we know from Corollary 3.11 that (22) and (23) can have arbitrarily small linear L^2 gains from $\tilde{\Delta}_1$ to $\tilde{P}_1$ and $\tilde{\Delta}_2$ to $\tilde{P}_2$; i.e., C_2/C_1 and C_3/C_1 can be made sufficiently small, on any compact sets, by choosing Q_1 and Q_2 properly. Then for any $\varepsilon > 0$, we can find Q_1 and Q_2, such that

$$\begin{aligned}
\dot{V}(\tilde{P}_1, \tilde{P}_2) \leq & -C_1(|\tilde{P}_1|^2 + |\tilde{P}_2|^2) + C_2|\tilde{P}_1|\gamma_{1,1}(|\tilde{P}_1|) + \frac{C_2}{2}|\tilde{P}_1|^2 + \frac{C_2}{2}\gamma_{1,2}^2(|\tilde{P}_2|) \\
& + \frac{C_1}{2}|\tilde{P}_2|^2 + \frac{C_3^2}{2C_1}\gamma_{2,1}^2(|\tilde{P}_1|) + C_3|\tilde{P}_2|\gamma_{2,2}(|\tilde{P}_2|) \\
\leq & -C_5(|\tilde{P}_1|^2 + |\tilde{P}_2|^2) \quad \forall |\tilde{P}_1| < \varepsilon, \ |\tilde{P}_2| < \varepsilon,
\end{aligned}$$

for some $C_5 > 0$. This completes the proof. □

Based on Theorem 4.8, we develop a coupled VI algorithm in Algorithm 4. The convergence of Algorithm 4 is given in the following theorem.

THEOREM 4.9. *Under Assumption* 4.6*, suppose that B_1 and B_2 have full rank. If $\sup_k\{h_{i,k}\}$ is sufficiently small, then given $Q_{i,0} \in \mathcal{S}_+^{n_i}$ and $R_{i,0} \in \mathcal{S}_+^{m_i}$, for any $\varepsilon > 0$, there exist $\lambda_i > 0$, such that by selecting $Q_i = \lambda_i Q_{i,0}$ and $R_i = R_{i,0}$, we have $\lim_{k\to\infty} P_{i,k} = P_i^*$, where $\{P_{i,k}\}_{k=0}^{\infty}$ is obtained from Algorithm* 4 *with $P_{i,0} \in \mathcal{S}^{n_i} \cap B_\varepsilon(P_i^*)$, and $i = 1, 2$.*

Proof. First, we show $\{P_{i,k}\}$ is bounded in $\mathcal{S}^{n_i} \cap B_\varepsilon(P_i^*)$. By picking λ_i sufficiently large, we know from part (i) of Corollary 3.13 that the coupled system (22)–(23) can be made asymptotically stable at (P_1^*, P_2^*), with $P_i(0) \in \mathcal{S}^{n_i} \cap B_\varepsilon(P_i^*)$, and we also know from Corollary 3.11 that ε can be made arbitrarily large.

Now, choosing $\sup_k\{h_{i,k}\}$ sufficiently small, we easily have from part (i) of Theorem 3.14 that $\{P_{i,k}\}$ stays in $\mathcal{S}^{n_i} \cap B_\varepsilon(P_i^*)$. Then, the proof is completed by Theorem 4.8 and following the proof of Theorem 3.14, part (iii). □

5. Illustrative practical examples.

In this section, we provide two simulation examples to illustrate our robust VI algorithm.

5.1. Mean-variance portfolio optimization based on multiplayer non-zero-sum differential game.

In this example, we study the mean-variance portfolio optimization problem [61] using non-zero-sum differential game theory obtained in subsection 4.2.

Consider the price process of $N + 1$ assets (or securities) traded continuously in a market [61]:

$$dS_0 = rS_0 dt,$$
$$dS_i = b_i S_i dt + \sum_{j=1}^{n_i} \sigma_{ij} S_i dw_j, \quad i = 1, 2, \dots, N,$$

where S_0 represents the price of a bond, S_i, $i = 1, \dots, N$, represent N stocks, $r > 0$ is the interest rate, $b_i > 0$ is the appreciation rate, and $\{\sigma_{ij}\}_{j=1}^{n_i}$ is the volatility of the ith stock. An investor's total wealth at time t, when holding $h_i(t)$ shares of the ith asset, is given as $x(t) = \sum_{i=0}^{N} h_i(t) S_i(t)$. The design objective here is to find h_i to (a) maximize the average return, and (b) minimize the volatility of x.

Inspired by [61], instead of solving the above portfolio optimization problem directly, we consider an auxiliary multiplayer non-zero-sum differential game composed with the following cost:

$$J_i(u) = \mathbb{E}\left[\int_0^\infty \left(Q_i \bar{x}^2 + \sum_{j=1}^{N} R_{ij} \bar{u}_i \bar{u}_j\right) dt\right], \quad i = 1, \dots N, \tag{24}$$

subject to

$$d\bar{x} = \left(r\bar{x} + \sum_i (b_i - r)\bar{u}_i\right) dt + \sum_{i,j} \sigma_{ij} \bar{u}_i dw_j,$$

where $\bar{x} = x - \gamma$, and $\gamma > 0$ represents the tradeoff between the two objectives in the portfolio optimization problem. A larger γ means more weights on the average return, and a small γ means more weights on the volatility. Note that the first term in the integrand in (24) is related to the variance of $\bar{x}$ (and hence x) at the steady state, and the second term guarantees that the shares for the ith bond/stock do not diverge to the infinity.

Since the volatilities of assets are usually difficult to estimate, we borrow the idea of stochastic robust optimal solution from [14, section 5], by choosing sufficiently small $Q_i > 0$ and $R_{ij} > 0$ to guarantee the small-gain condition. Then, the above non-zero-sum differential game can be solved using Algorithm 4, with $A_i = r$ and $B_i = b_i - r$. Based on the desired expected return, γ is chosen as 200. Once $\bar{u}_i^* := -K_i^* \bar{x}$ is obtained, the optimal share of the ith asset at time t is chosen as $K_i^*(\gamma - \bar{x}(t))$. In total 20 stocks and one bond are used to construct the portfolio. The interest rate is chosen as 2.5%, and the appreciation rates are randomly selected from 0% to 15%. After 1000 iterations, all P_i's converge to their optimal values. The prices of the portfolio and the paths of P_i's are shown in Figure 1. Note that the portfolio constructed using the non-zero-sum differential game approach has a higher return, while maintaining approximately the same volatility compared with the uniform allocation of the asset.

5.2. ADP for time-series variance minimization. In this example, we use the ADP method developed in subsection 4.1 to study the variance minimization problem for a class of time series with unknown parameters. Note that this is a classical problem which has been studied in both finance and the signal processing community, and can be easily addressed using the Kalman filter when the model parameters are known.

Consider the following time series in continuous time:

$$\dddot{S} = \alpha_3 \ddot{S} + \alpha_2 \dot{S} + \alpha_1 S + \sigma_0 v_0,$$

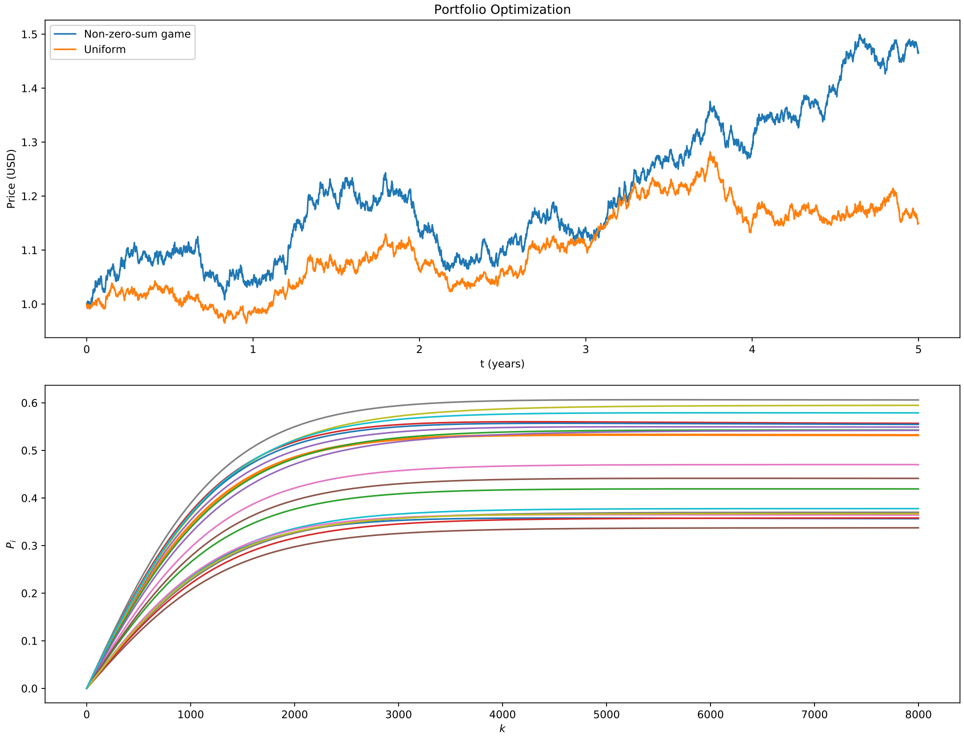

FIG. 1. *Example* 5.1: *Portfolio optimization solved from robust VI.*

where σ_0 and α_i, $i = 1, 2, 3$, are unknown model parameters, and v_0 is a Gaussian white noise that drives the output S. Suppose the system is asymptotically stable in mean square sense. Our objective here is to minimize the variance of S.

We can rewrite the above differential equation in state space form, by assuming states $x_1 = S + \sigma_1 w_1$, $x_2 = \dot{S} + \sigma_2 w_2$, $x_3 = \ddot{S} + \sigma_3 w_3$ and a control input u. Here w_i, $i = 0, 1, 2, 3$, are Brownian motions representing the observation noises, and σ_i, $i = 1, 2, 3$, are unknown noise magnitudes. Note that even if $u \equiv 0$, $\mathbb{E}x_i$, $i = 1, 2, 3$, can decrease to 0 asymptotically since we assume the system is asymptotically stable in mean square sense. However, the variance of x_i may be extremely large due to the presence of $\sigma_0 v_0$. To reduce the variance of x_i, Algorithm 3 is used to develop an ergodic controller. Notice that by the law of large numbers, $\int_0^\infty w_i dt = 0$ for all i. Hence, the two terms $\sigma_2 w_2 dt$ and $\sigma_3 w_3 dt$ have little influence in the time integration in Algorithm 3.

In the simulation, we choose $\alpha_1 = -4$, $\alpha_2 = -1$, $\alpha_3 = -4$, $\sigma_0 = 1$, $\sigma_1 = 0.6$, $\sigma_2 = 0.4$, and $\sigma_3 = 0.5$. For illustration purposes, the weighting matrices in the cost are chosen as $Q = 0.1I$ and $R = 0.01$. P_k is updated in real time after every second. Both the optimal solution P^* and the near-optimal solution $\hat{P}^*$ from ADP learning are shown below:

$$P^* = \begin{bmatrix} 0.2859 & 0.1492 & 0.0110 \\ 0.1492 & 0.3366 & 0.0539 \\ 0.0110 & 0.0539 & 0.0206 \end{bmatrix}, \quad \hat{P}^* = \begin{bmatrix} 0.2854 & 0.1479 & 0.0106 \\ 0.1479 & 0.3377 & 0.0529 \\ 0.0106 & 0.0529 & 0.0262 \end{bmatrix}.$$

The system trajectories and P_k are given in Figure 2. Note that the controller derived

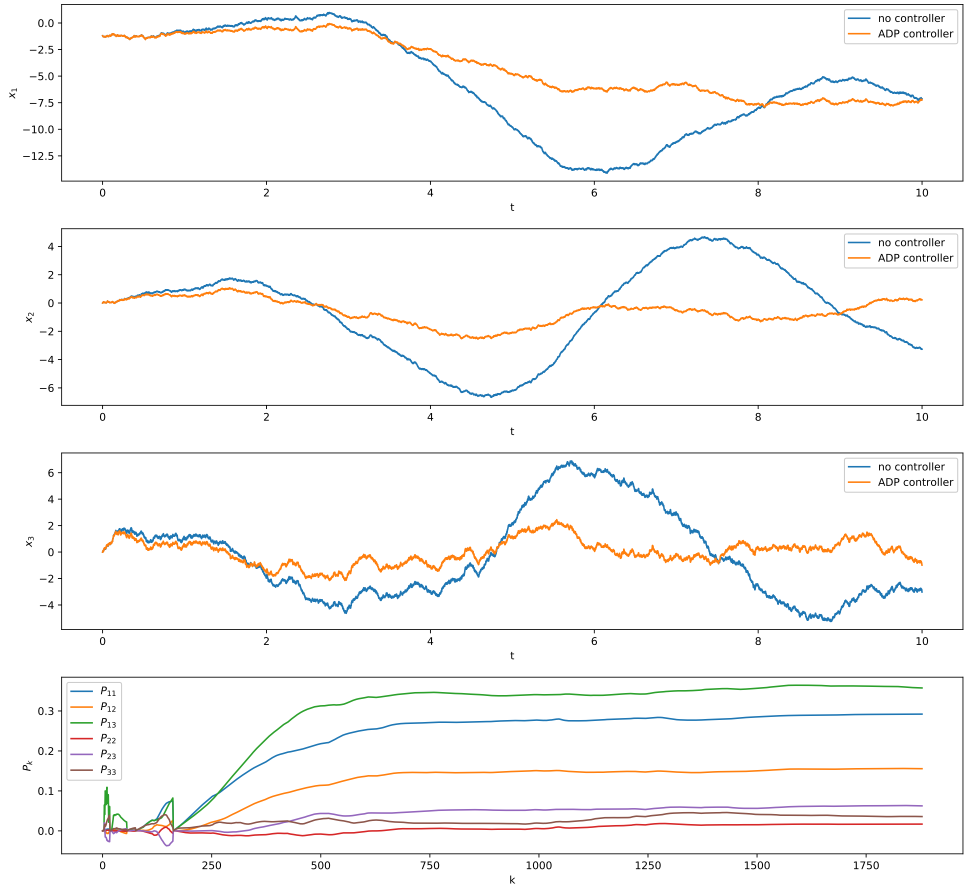

FIG. 2. *Example* 5.2: *The trajectories of* x_i, $i = 1, 2, 3$, *and* P_k.

from Algorithm 3 significantly reduces the variance of the output signal.

6. Summary.

This paper develops a new framework of robust DP. This novel theory resolves a long-standing issue in DP theory: how to develop DP algorithms that are robust to different types of disturbances. Empowered by nonlinear and robust control theories, robust DP allows us to develop various DP and RL algorithms with guaranteed convergence to the optimal solution in the presence of different types of disturbances, including stochastic noise, external disturbances, and modeling errors such as nonlinear dynamic uncertainties. To be specific, we have conducted an innovative input-output gain analysis for the DMRE in section 3, and applied the result together with the nonlinear small-gain theory to develop a novel robust VI algorithm. It has been shown that this new algorithm is robust to different kinds of internal and external disturbances, and hence is especially useful in solving non-model-based optimal control problems.

Due to space limitation, we list only a few illustrative applications of our robust DP method in section 4. These examples have demonstrated that robust DP obtained in the present paper is a powerful tool for addressing adaptive optimal control and DP problems.

Appendix A. Proofs of some technical results.

A.1. Proof of Theorem 3.8. To prove part (i), we first introduce the following finite-horizon cost:

$$\mathcal{J}_t(x(t); u, Q_\Delta) = x^T(0)P_\Delta(0)x(0) + \int_t^0 (x^T(s)Q_\Delta(s)x(s) + u^T(s)Ru(s))ds,$$

where $t < 0$ is an arbitrary time instant, and $Q_\Delta(s) = Q + \Delta(s)$. Since $Q_\Delta(s) \geq 0$ on $[t, 0]$, we know from the linear quadratic regulator theory [36, Chapter 6.1] that $\inf_u \mathcal{J}_t(x(t); u, Q_\Delta) = x^T(t)M(t)x(t)$, where $M(s) = M^T(s) > 0$, $s \in [t, 0]$, satisfies

$$-\dot{M} = A^T M + MA - MBR^{-1}B^T M + Q + \Delta, \quad M(0) = P_\Delta(0).$$

Moreover, the optimal controller for $\mathcal{J}_t$ is $u^o(s) := -R^{-1}B^T M(s)$.

Next, we have from the conditions on Q_Δ that there exists a constant matrix $\overline{Q} \in \mathcal{S}$, such that $0 \leq Q_\Delta(s) < \overline{Q}$ for all s. Thus,

$$\begin{aligned} 0 \leq x^T(t)M(t)x(t) &\leq x^T(0)P_\Delta(0)x(0) + \int_t^0 (x^T Q_\Delta(s)x + (\bar{u}^o)^T R\bar{u}^o)ds \\ &\leq x^T(0)P_\Delta(0)x(0) + \int_t^0 (x^T\overline{Q}x + (\bar{u}^o)^T R\bar{u}^o)ds, \end{aligned} \tag{25}$$

where $\bar{u}^o := \arg\inf_u \mathcal{J}_t(x(t); u, \overline{Q})$. Since $\overline{Q}$ is positive definite, we know there exists a real symmetric matrix $\overline{M} > 0$, such that

$$\inf_u \mathcal{J}_t(x(t); u, \overline{Q}) < x^T(t)\overline{M}x(t).$$

Then, we have from (25) that $0 \leq M(t) < \overline{M}$ for all $t < 0$. Comparing the definitions of M and P_Δ, we know $M(t) = P_\Delta(-t)$. Thus, $0 \leq P_\Delta(t) < \overline{M}$ for all $t > 0$.

To prove part (ii), note from part (i) that P_Δ is bounded on $\mathbb{R}_+$. Then, since $P(t)$ converges to P^*, for any $\varepsilon > 0$, there exists $T_0 > 0$, such that $\sup_{T>T_0} |P(t+T)-P^*| < \varepsilon$, given $P(t) = P_\Delta(t)$ for any $t > 0$. In addition, by [45, Theorem 55], for any $T_1 > 0$ and $\varepsilon > 0$, we can find $t_0 > 0$ under which $\sup_{t\geq t_0} |\Delta(t)|$ is sufficiently small, so that $\sup_{T\in[0,T_1]} |P(t + T) - P_\Delta(t + T)| < \varepsilon$, given $P(t) = P_\Delta(t)$ for all $t > t_0$.

Now, by picking $T_1 = 2T_0$, one can guarantee from the above analysis that $|P^* - P_\Delta(t + T)| < 2\varepsilon$ for all $t > t_0$ and $T \in [T_0, 2T_0]$. Thus, we know $\sup_{t>t_0+T_0} |P_\Delta(t) - P^*| \leq 2\varepsilon$. Since t_0 exists for any ε, which can be made arbitrarily small, we have $\lim_{t\to\infty} P_\Delta(t) = P^*$.

Moreover, choosing the same Lyapunov function in the proof of Proposition 3.6, we know there exist positive constants C_1, C_2, and ε_1, such that

$$\dot{V}(\tilde{P}_\Delta) \leq -C_1|\tilde{P}_\Delta|^2 + C_2|\tilde{P}_\Delta||\Delta| \quad \forall |\tilde{P}_\Delta| < \varepsilon_1, \tag{26}$$

where $\tilde{P}_\Delta = P_\Delta - P^*$. By completing the squares, we have from (26) that (10) admits a finite linear L^2 gain in a neighborhood of P^*. Thus, by H^∞ control theory [56], $\tilde{P}_\Delta \in L^2$ if $\Delta \in L^2$.

Now, we prove part (iii). Note from (12) and (26) that by defining $\bar{V}(P, M) = V(P) + \frac{C_2^2}{C_1}V_f(M)$,

$$\begin{aligned} \frac{d}{dt}\bar{V}(\tilde{P}_\Delta, \tilde{M}) &\leq -C_1|\tilde{P}_\Delta|^2 + C_2|\tilde{P}_\Delta||\Delta| - \frac{C_2^2}{C_1}|\Delta|^2 + \frac{C_2^2}{C_1}\gamma^2|\tilde{P}_\Delta|^2 \\ &= -\left(\frac{C_1}{2} - \frac{C_2^2}{C_1}\gamma^2\right)|\tilde{P}_\Delta|^2 - \frac{C_2^2}{2C_1}|\Delta|^2 \quad \forall |\tilde{P}_\Delta| < \varepsilon_1, \ |\tilde{M}| < \varepsilon_0. \end{aligned}$$

Take $\gamma < \frac{C_1}{\sqrt{2}C_2}$. Since (11) is zero-state detectable, we have from LaSalle's invariance principle [28, Corollary 4.1] that (P_Δ, M) is asymptotically stable at (P^*, M^*).

Finally, to prove part (iv) involving stochastic disturbance, from Itô's lemma [49] and (26), it follows that

$$\mathcal{L}V(\tilde{P}_\Delta) \leq -C_1|\tilde{P}_\Delta|^2 + C_3\sum_{i=1}^{N}|\Delta_i|^2 \quad \forall|\tilde{P}_\Delta| < \varepsilon,$$

for some positive constants ε and C_3, where $\mathcal{L}$ denotes the differential generator. Note that C_3 is bounded because $\partial_x^2 V$ is bounded on any compact sets, since V is smooth. Obviously, if we pick $\gamma = C_1/(2C_3)$, then

$$\mathcal{L}V(\tilde{P}_\Delta) \leq -\frac{1}{2}C_1|\tilde{P}_\Delta|^2 \quad \forall|\tilde{P}_\Delta| < \varepsilon.$$

This concludes the proof. □

A.2. Proof of Theorem 3.14. The proof is inspired by the analysis in [18]. Before proving part (i), we denote an operator $\mathcal{R} : \mathcal{S}^n \to \mathcal{S}^n$, such that

$$\mathcal{R}(P) = A^T P + PA - PBR^{-1}B^T P + Q.$$

Suppose $P_0 \neq P^*$. By Lemma 3.3 and shifting the equilibrium of (5) to P^*, we know there exists a smooth Lyapunov function $\mathcal{V} : R_A \to \mathbb{R}_+$, where $R_A \subset \mathcal{S}^n$ is the region of attraction of P^*, such that[8]

$$\begin{aligned}\langle \partial_x \mathcal{V}(P), \mathcal{R}(P)\rangle_F < 0, \quad \mathcal{V}(P) > 0 \quad \forall P \in R_A \setminus \{P^*\},\\ \lim_{P\to\partial R_A} \mathcal{V}(P) = \infty, \quad \langle \partial_x \mathcal{V}(P^*), \mathcal{R}(P^*)\rangle_F = 0 \quad \mathcal{V}(P^*) = 0.\end{aligned}$$

As a result, $\{P : \mathcal{V}(P) \leq C\}$ is a compact subset of R_A for all $C > 0$. Then, there exist $C_0 > 0$ and $C_1 > 0$, such that $C_0 < \mathcal{V}(P_0) < C_1$. Furthermore, we can find a sufficiently small constant $\varepsilon_\delta > 0$, such that for all $\zeta \in \mathcal{S}^n$ satisfying $|\zeta| < \varepsilon_\delta$,

$$\sup_{\{P:C_0\leq\mathcal{V}(P)\leq C_1\}} \{\langle \partial_x \mathcal{V}(P), (\mathcal{R}(P) + \zeta)\rangle_F\} = -\delta \tag{27}$$

for some $\delta > 0$.

By contradiction, suppose $\{P_k\}_{k=0}^\infty$ is unbounded. Then, there exists an upcrossing interval $[C_2, C_3]$, with $\mathcal{V}(P_0) < C_2 < C_3 < C_1$, such that $\{\mathcal{V}(P_k)\}_{k=0}^\infty$ crosses this interval from below infinitely many times.

From the conditions on W_k, we know there exists $E \in \mathcal{F}$ with $\mathbb{P}(E) = 1$, such that for all $\omega \in E$, $\{W_k(\omega)\}_{k=0}^\infty$ is bounded. Fixing $\omega \in E$, we can define two subsequences $\{P_{k_j}\}, \{P_{k'_j}\} \subset \{P_k\}$, such that

$$\mathcal{V}(P_{k_j-1}) < C_2 \leq \mathcal{V}(P_m) < C_3 < \mathcal{V}(P_{k'_j}) \quad \forall k_j \leq m < k'_j. \tag{28}$$

Choose a sufficiently small $\varepsilon > 0$, such that for any $P \in \{P_{k_j}\}$, $B_\varepsilon(P) \subset \{P \in \mathcal{S}_+^n :$

[8] Note that $\mathcal{V}$ defined here is different from the Lyapunov function used in the proof of Proposition 3.6.

$\mathcal{V}(P) < C_1\}$. Suppose q is sufficiently large. Then, for any $j \in \mathbb{Z}_+$,

$$\begin{aligned}\varepsilon < |P_{L_\varepsilon(j)} - P_{k_j}| &= \left| \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i(\mathcal{R}(P_i) + \Delta_i + W_i) \right| \\ &\le \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i(|\mathcal{R}(P_i)| + |\Delta_i| + |W_i|) \le \varepsilon_C \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i,\end{aligned}$$

where $L_\varepsilon(j) = \inf\{i \ge k_j : |P_i - P_{k_j}| > \varepsilon\}$, and $\varepsilon_C > 0$ is a constant independent of j.

Then, by the assumption on W_k, one has

$$\begin{aligned}&\mathcal{V}(P_{L_\varepsilon(j)}) - \mathcal{V}(P_{k_j}) \\ &= \int_0^1 \left\langle \partial_x \mathcal{V}(P_{k_j} + t(P_{L_\varepsilon(j)} - P_{k_j})), (P_{L_\varepsilon(j)} - P_{k_j}) \right\rangle_F dt \\ &= \left\langle \partial_x \mathcal{V}(P_{k_j}), (P_{L_\varepsilon(j)} - P_{k_j}) \right\rangle_F \\ &\quad + \int_0^1 \int_0^1 \left\langle \frac{d}{ds} \partial_x \mathcal{V}(P_{k_j} + st(P_{L_\varepsilon(j)} - P_{k_j})), (P_{L_\varepsilon(j)} - P_{k_j}) \right\rangle_F dsdt \\ &= \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i \left\langle \partial_x \mathcal{V}(P_{k_j}), (\mathcal{R}(P_{k_j}) + \bar{\Delta}_{i,j}) \right\rangle_F + \left\langle \partial_x \mathcal{V}(P_{k_j}), \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i W_i \right\rangle_F \\ &\quad + \int_0^1 \int_0^1 \left\langle \frac{d}{ds} \partial_x \mathcal{V}(P_{k_j} + st(P_{L_\varepsilon(j)} - P_{k_j})), (P_{L_\varepsilon(j)} - P_{k_j}) \right\rangle_F dsdt, \qquad (29)\end{aligned}$$

where $\bar{\Delta}_{i,j} = \Delta_i + \mathcal{R}(P_i) - \mathcal{R}(P_{k_j})$. Note that $\lim_{j\to\infty} |P_{L_\varepsilon(j)} - P_{k_j}| = \varepsilon$, because $\lim_{k\to\infty} h_k = 0$. Then, since P_{k_j} is bounded and $\mathcal{V}$ is smooth,

$$\lim_{j\to\infty} \left| \int_0^1 \int_0^1 \left\langle \frac{d}{ds} \partial_x \mathcal{V}(P_{k_j} + st(P_{L_\varepsilon(j)} - P_{k_j})), (P_{L_\varepsilon(j)} - P_{k_j}) \right\rangle_F dsdt \right| = O(\varepsilon^2).$$

Since $\lim_{j\to\infty} \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i W_i = 0$, there exists a sufficiently large $\bar{j}$, such that for all $j > \bar{j}$, by choosing sufficiently small ε and δ_0, we have $|\bar{\Delta}_{i,j}| < \varepsilon_\delta$, and by (27) and (29) it follows that

$$\begin{aligned}\mathcal{V}(P_{L_\varepsilon(j)}) - \mathcal{V}(P_{k_j}) &\le \left\langle \partial_x \mathcal{V}(P_{k_j}), \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i W_i \right\rangle_F - \delta \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i + O(\varepsilon^2) \\ &\le \left\langle \partial_x \mathcal{V}(P_{k_j}), \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i W_i \right\rangle_F - \frac{\delta\varepsilon}{\varepsilon_C} + O(\varepsilon^2) < 0.\end{aligned}$$

Since $\lim_{k\to\infty} \mathcal{V}(P_{k_j}) = C_2$, we know that for a large enough k, if $P_k \in \{P_{k_j}\}$, then there exists $k' > k$, such that $\mathcal{V}(P_{k'}) < C_2$, and P_i stays in a ε-neighborhood of P_k for $k \le i \le k'$. Thus, P_k is bounded, and the proof of part (i) is concluded by contradiction.

Now, we prove part (ii). First, rewrite the updating equation in Algorithm 2 as

$$P_{k+1} = P_k + h_k(\mathcal{R}(P_k) + \Delta_k + W_k) + Z_k, \qquad k \ge N,\ P_k \in \mathcal{S}_0,$$

where N is chosen as in part (i), and the projection term Z_k is defined as

$$Z_k = \begin{cases} P_0 - P_{k+1/2} & \text{if } P_{k+1/2} \notin \mathcal{S}_0, \\ 0 & \text{otherwise.} \end{cases}$$

Define the following continuous-time interpolation:

$$P^0(t) = \begin{cases} P_0, & t \leq 0, \\ P_k, & t \in [t_k, t_{k+1}), \end{cases} \qquad \Delta^0(t) = \begin{cases} \Delta_0, & t \leq 0, \\ \Delta_k, & t \in [t_k, t_{k+1}), \end{cases}$$

where $t_0 = 0$ and $t_k = \sum_{i=0}^{k-1} h_i$ for $k \geq 1$. Define the shifted process $P^k(t) = P^0(t_k+t)$ and $\Delta^k(t) = \Delta^0(t_k + t)$ for all $t \in \mathbb{R}$.

Then, we have for all $k \geq N$ and $t \geq 0$ that

$$\begin{aligned} P^k(t) &= P_k + \sum_{i=k}^{m(t+t_k)-1} h_i(\mathcal{R}(P_i) + \Delta_i) + W^k(t) + Z^k(t) \\ &= P_k + H^k(t) + e^k(t) + W^k(t) + Z^k(t), \end{aligned} \tag{30}$$

where

$$H^k(t) = \int_0^t (\mathcal{R}(P^k(s)) + \Delta^k(s))ds, \quad Z^k(t) = \sum_{i=k}^{m(t+t_k)-1} Z_i,$$

$$W^k(t) = \sum_{i=k}^{m(t+t_k)-1} h_i W_i, \quad m(t) = \begin{cases} j, & 0 \leq t_j \leq t < t_{j+1}, \\ 0, & t < 0, \end{cases}$$

and $e^k(t)$ is due to replacing $\sum_{i=k}^{m(t+t_k)-1} h_i(\mathcal{R}(P_i) + \Delta_i)$ with $H^k(t)$. By convention, the above definition assumes $\sum_{i=k}^{m(t+t_k)-1} * = 0$, when $0 \leq t < h_k$. Note that for all $\omega \in E$, $W^k(\cdot, \omega)$ converges to 0 uniformly on any finite time interval.

Fixing $T > 0$ and following the proof of [14, Theorem 3.3], we can show that $\{H^k(\cdot)\}_{k=N}^\infty$, $\{Z^k(\cdot)\}_{k=N}^\infty$, and $\{e^k(\cdot)\}_{k=N}^\infty$ are all relatively compact in $\mathcal{D}([0, T], \mathcal{S}^n)$, where $\mathcal{D}([0, T], \mathcal{S}^n)$ denotes the space of functions from $[0, T]$ to $\mathcal{S}^n$ that are right-continuous with left-hand limits, equipped with the Skorokhod topology [44]. Following the procedure in the proof of [1, Lemma 3.4], one can show that the limit of $\{Z^k(\cdot)\}_{k=N}^\infty$ is identically 0. Then, the limit of $\{P_k, \Delta_k\}$ satisfies

$$\dot{P} = \mathcal{R}(P) + \Delta,$$

where Δ converges to 0 by its definition. By part (i), we know $\{P_k\}_{k=N}^\infty$ remains in the region of attraction of P^*. Thus, part (ii) is established by Theorem 3.8, part (ii), and Part 2 of the proof of [33, Theorem 5.2.1].

To prove part (iii), we note from part (iii) of Theorem 3.8 that the following coupled system is asymptotically stable at (P^*, M^*):

$$\begin{aligned} \dot{P} &= \mathcal{R}(P) + \Delta(P, M), \\ \dot{M} &= f(M, P). \end{aligned}$$

Moreover, by defining $\bar{\mathcal{V}}(P, M) = \bar{V}(P - P^*, M - M^*)$, where the Lyapunov function $\bar{V}$ is defined in the proof of Theorem 3.8, we also have

$$\langle \partial_P \bar{\mathcal{V}}(P, M), (\mathcal{R}(P) + \Delta) \rangle_F + \langle \partial_M \bar{\mathcal{V}}(P, M), f(M, P) \rangle_F < 0$$

for all (P, M) in a small neighborhood of (P^*, M^*) with $(P, M) \neq (P^*, M^*)$. Since M_k is bounded, Δ_k is bounded for all bounded P_k. Now, following the steps in part (i), we can show (P_k, M_k) is bounded, provided P_0 stays in a small neighborhood of P^*, and ε_0 is small enough. Applying the analysis in part (ii), we know (P_k, M_k) converges to the solution to the above coupled ODE. By part (iii) of Theorem 3.8, this completes the proof. □

A.3. Proof of Corollary 3.17. We need only show that P_k is bounded. Then, we easily have $\sum_{k=0}^{\infty} h_k W_k < \infty$ with probability one [18, Remark 1], and the convergence is proved by part (ii) of Theorem 3.14.

By contradiction, suppose $\{P_k\}_{k=0}^{\infty}$ is unbounded. Following the analysis in the proof of Theorem 3.14, part (i), we still have

$$\varepsilon < |P_{L_\varepsilon(j)} - P_{k_j}| = \left| \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i(\mathcal{R}(P_i) + \Delta_i(P_i) + \sigma_i(P_i)v_i) \right| \leq \varepsilon_C \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i$$

for some $\varepsilon_C > 0$, where ε, k_j, $L_\varepsilon(j)$, and $\mathcal{R}$ follow the same definitions as in Appendix A.2. Since $\lim_{k\to\infty} \Delta_k = 0$ uniformly on any compact set, $\sup_{i\in[k_j, L_\varepsilon(j)]\cap\mathbb{Z}_+} |\Delta_i(P_i)|$ can be made arbitrarily small by choosing a large enough j. Then, there exists a sufficiently large $\bar{j}$, such that for all $j > \bar{j}$,

$$V(P_{L_\varepsilon(j)}) - V(P_{k_j}) \leq \left\langle \partial_x V(P_{k_j}), \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i \sigma_i(P_i) v_i \right\rangle_F - \frac{\delta\varepsilon}{\varepsilon_C} + O(\varepsilon^2). \tag{31}$$

Now, define a sequence $\{M_k\}$, such that

$$M_k = \sum_{i \in \cup_{j\in\{j\in\mathbb{Z}_+ : k_j' \leq k\}} ([k_j, L_\varepsilon(j)-1]\cap\mathbb{Z}_+)} h_i \sigma_i(P_i) v_i,$$

where k_j' is defined in (28). Obviously, $\{M_k\}$ is a martingale with respect to $\{\mathcal{F}_k\}$, and $\mathbb{E}\left[|M_k|^2\right]$ is bounded, since $\{P_i\}_{i\in[k_j, L_\varepsilon(j)]\cap\mathbb{Z}_+}$ is bounded, $\sum_{k=0}^{\infty} h_k^2 < \infty$, and v_k has finite variance. By the martingale convergence theorem [49, Theorem 2.6], M_k converges with probability one, and thus $\lim_{j\to\infty} \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i \sigma_i(P_i) v_i = 0$. This, together with (31), shows that P_k is bounded with probability one. □

REFERENCES

[1] J. Abounadi, D. P. Bertsekas, and V. Borkar, *Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms*, SIAM J. Control Optim., 41 (2002), pp. 1–22, https://doi.org/10.1137/S0363012998346621.

[2] A. Arapostathis, V. S. Borkar, and M. K. Ghosh, *Ergodic Control of Diffusion Processes*, Cambridge University Press, New York, 2012.

[3] R. W. Beard, *Improving the Closed-Loop Performance of Nonlinear Systems*, Ph.D. thesis, Rensselaer Polytechnic Institute, 1995.

[4] R. Bellman, *On the theory of dynamic programming*, Proc. Nat. Acad. Sci. U.S.A., 38 (1952), pp. 716–719.

[5] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[6] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. 1, 3rd ed., Athena Scientific, Belmont, MA, 2005.

[7] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. 2, 3rd ed., Athena Scientific, Belmont, MA, 2007.

[8] D. P. Bertsekas, *Abstract Dynamic Programming*, Athena Scientific, Belmont, MA, 2013.

[9] D. P. Bertsekas, *Value and policy iterations in optimal control and adaptive dynamic programming*, IEEE Trans. Neural Netw. Learn. Syst., 28 (2017), pp. 500–509.

[10] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[11] T. Bian, Y. Jiang, and Z.-P. Jiang, *Adaptive dynamic programming and optimal control of nonlinear nonaffine systems*, Automatica J. IFAC, 50 (2014), pp. 2624–2632.

[12] T. Bian and Z.-P. Jiang, *Stochastic adaptive dynamic programming for robust optimal control design*, in Control of Complex Systems: Theory and Applications, K. G. Vamvoudakis and S. Jagannathan, eds., Butterworth-Heinemann, Cambridge, MA, 2016, pp. 211–245.

[13] T. Bian and Z.-P. Jiang, *Value iteration, adaptive dynamic programming, and optimal control of nonlinear systems*, in Proceedings of the 2016 IEEE 55th Conference on Decision and Control (CDC), Las Vegas, 2016, pp. 3375–3380.

[14] T. Bian and Z.-P. Jiang, *Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design*, Automatica J. IFAC, 71 (2016), pp. 348–360.

[15] V. S. Borkar, *Ergodic control of diffusion processes*, in Proceedings of the International Congress of Mathematicians, Vol. III, 2006, pp. 1299–1309.

[16] J. W. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits Syst., 25 (1978), pp. 772–781.

[17] D. L. Burkholder, B. J. Davis, and R. F. Gundy, *Integral inequalities for convex functions of operators on martingales*, in Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 2, Berkeley, CA, 1972, pp. 223–240.

[18] H.-F. Chen and Y. Zhu, *Stochastic approximation procedures with randomly varying truncations*, Sci. Sinica Ser. A, 29 (1986), pp. 914–926.

[19] M. Chilali, P. Gahinet, and P. Apkarian, *Robust pole placement in LMI regions*, IEEE Trans. Automat. Control, 44 (1999), pp. 2257–2270.

[20] R. A. Howard, *Dynamic Programming and Markov Processes*, The MIT Press, Cambridge, MA, 1960.

[21] A. Isidori, *Nonlinear Control Systems* II, Springer, London, 1999.

[22] G. N. Iyengar, *Robust dynamic programming*, Math. Oper. Res., 30 (2005), pp. 257–280.

[23] Y. Jiang and Z.-P. Jiang, *Robust adaptive dynamic programming and feedback stabilization of nonlinear systems*, IEEE Trans. Neural Netw. Learn. Syst., 25 (2014), pp. 882–893.

[24] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*, Wiley-IEEE Press, Hoboken, NJ, 2017.

[25] Z.-P. Jiang and Y. Jiang, *Robust adaptive dynamic programming for linear and nonlinear systems: An overview*, Eur. J. Control, 19 (2013), pp. 417–425.

[26] Z.-P. Jiang and T. Liu, *Small-gain theory for stability and control of dynamical networks: A survey*, Annu. Rev. Control, 46 (2018), pp. 58–79.

[27] Z.-P. Jiang, A. R. Teel, and L. Praly, *Small-gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1994), pp. 95–120.

[28] H. K. Khalil, *Nonlinear Systems*, 3rd ed., Prentice Hall, Upper Saddle River, NJ, 2002.

[29] R. Khas'minskii, *Stochastic Stability of Differential Equations*, Springer, Berlin, Heidelberg, 2012.

[30] D. L. Kleinman, *On an iterative technique for Riccati equation computations*, IEEE Trans. Automat. Control, 13 (1968), pp. 114–115.

[31] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley & Sons, New York, 1995.

[32] V. Kučera, *A review of the matrix Riccati equation*, Kybernetika (Prague), 9 (1973), pp. 42–61.

[33] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, New York, 2003.

[34] H. Kwakernaak and R. Sivan, *The maximally achievable accuracy of linear optimal regulators and linear optimal filters*, IEEE Trans. Automat. Control, 17 (1972), pp. 79–86.

[35] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, John Wiley & Sons, Piscataway, NJ, 2013.

[36] D. Liberzon, *Calculus of Variations and Optimal Control Theory: A Concise Introduction*, Princeton University Press, Princeton, NJ, 2012.

[37] S. H. Lim, H. Xu, and S. Mannor, *Reinforcement learning in robust Markov decision processes*, in Advances in Neural Information Processing Systems 26, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds., Curran Associates, 2013, pp. 701–709.

[38] T. Liu, Z.-P. Jiang, and D. J. Hill, *Nonlinear Control of Dynamic Networks*, CRC Press, Boca Raton, FL, 2014.

[39] A. Nilim and L. El Ghaoui, *Robust control of Markov decision processes with uncertain transition matrices*, Oper. Res., 53 (2005), pp. 780–798.

[40] H. Pham, *Continuous-time Stochastic Control and Optimization with Financial Applications*, Springer-Verlag, Berlin, Heidelberg, 2009.

[41] L. Praly and Y. Wang, *Stabilization in spite of matched unmodeled dynamics and an equivalent definition of input-to-state stability*, Math. Control Signals Systems, 9 (1996), pp. 1–33.

[42] M. Reed and B. Simon, *Methods of Modern Mathematical Physics: Functional Analysis*, Academic Press, San Diego, 1980.

[43] S. Sastry, *Nonlinear Systems: Analysis, Stability, and Control*, Springer-Verlag, New York, 1999.

[44] A. V. Skorokhod, *Limit theorems for stochastic processes*, Theory Probab. Appl., 1 (1956), pp. 261–290.

[45] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, 2nd ed., Springer, New York, 1998.

[46] E. D. Sontag, *Input to state stability: Basic concepts and results*, in Nonlinear and Optimal Control Theory: Lectures given at the C.I.M.E. Summer School held in Cetraro, Italy, June 19-29, 2004, P. Nistri and G. Stefani, eds., Springer-Verlag, Berlin, Heidelberg, 2008, pp. 163–220.

[47] A. W. Starr and Y. C. Ho, *Nonzero-sum differential games*, J. Optim. Theory Appl., 3 (1969), pp. 184–206.

[48] A. W. Starr and Y. C. Ho, *Further properties of nonzero-sum differential games*, J. Optim. Theory Appl., 3 (1969), pp. 207–219.

[49] J. M. Steele, *Stochastic Calculus and Financial Applications*, Springer, New York, 2001.

[50] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., The MIT Press, Cambridge, MA, 2018.

[51] R. S. Sutton, A. G. Barto, and R. J. Williams, *Reinforcement learning is direct adaptive optimal control*, IEEE Control Syst. Mag., 12 (1992), pp. 19–22.

[52] G. Tao, *Adaptive Control Design and Analysis*, John Wiley & Sons, Hoboken, NJ, 2003.

[53] E. Todorov, *Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system*, Neural Comput., 17 (2005), pp. 1084–1108.

[54] J. N. Tsitsiklis, *Asynchronous stochastic approximation and Q-learning*, Machine Learning, 16 (1994), pp. 185–202.

[55] J. N. Tsitsiklis and B. Van Roy, *An analysis of temporal-difference learning with function approximation*, IEEE Trans. Automat. Control, 42 (1997), pp. 674–690.

[56] A. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, 3rd ed., Springer International, Cham, 2017.

[57] D. Wang, H. He, and D. Liu, *Adaptive critic nonlinear robust control: A survey*, IEEE Trans. Cybernet., 47 (2017), pp. 3429–3451.

[58] D. Wang and D. Liu, *Neural robust stabilization via event-triggering mechanism and adaptive learning technique*, Neural Networks, 102 (2018), pp. 27–35.

[59] J. L. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[60] G. Zames, *On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity*, IEEE Trans. Automat. Control, 11 (1966), pp. 228–238.

[61] X. Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.