

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, four illustrative applications in the fields of stochastic optimal control and adaptive DP are presented. Three numerical examples arising from both finance and engineering industries are also given, along with several possible extensions of the proposed framework.

CONTENTS

1. Introduction	1
2. Preliminaries	4
2.1. System description	4
2.2. DMRE and continuous-time VI	4
3. Robust DP and VI for continuous-time systems	5
3.1. Robust DP and DMRE	5
3.2. Robust VI algorithm	12
4. Applications to adaptive/stochastic/decentralized optimal control	16
4.1. VI in the presence of modeling errors	17
4.2. VI-based ADP for linear continuous-time systems	18
4.3. Stochastic ADP for ergodic control problems	21
4.4. Decentralized VI	24
5. Illustrative practical examples	26
5.1. Mean-variance portfolio optimization	26
5.2. ADP learning for kinematic models	27
5.3. ADP for time-series variance minimization	29
6. Summary and future work	30
Acknowledgements	33
References	34

1. INTRODUCTION

In 1952, Bellman proposed the original idea of dynamic programming (DP) [Bel52] to solve a class of optimization problems subject to a controlled process

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.

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 [Put05, Ber05, Ber07, Ber13] have attracted a significant amount of attention, because of the vital role they have played in several popular fields including reinforcement learning (RL) [SB98, Lit15, BTS17], finance [Mer71, TVR99, Pha09, GP16], and biological control [KBL70, Tod05], to name a few. Depending on the form (discrete-time vs. 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, and numerous methods including policy iteration (PI) [How60, Kle68, Bea95, BJJ14] and value iteration (VI) [Bel57, Ber17, BJ16b, BJ16a] 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 disturbance in practical problems. Nevertheless, from a control theory point of view, we identify two perspectives to address these issues. The first one, 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 component can be the Q-factor, the policy gradient, and the policy function. Indeed, the majority of existing adaptive optimal control and DP methods [BT96, SBPW04, Pow07, LL13, JJ17] falls into this category, and RL is considered as a machine learning reinterpretation of direct adaptive control [SBW92]. The main advantage of these methods is that they are effective in tackling the presence of 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 [LJH14] 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 to the DP 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. 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 disturbances.

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 methods is that it is effective in dealing with both static and dynamic uncertainties. Unlike the adaptive control perspective, the development in this direction is still rudimentary. Only a few results [Iye05, NEG05, LXM13] are available for solving DP and RL problems along this track, in which the authors still only considered the static uncertainty caused by the unknown transition probability measures. Besides, these methods are only available for MDPs. In other words, there is no robust DP solution for dynamical

TABLE 1. Comparison between Applications 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

systems described by differential equations. As a result, it is still an open problem how to develop DP algorithms that are robust to both static and dynamic uncertainties.

In this paper, we propose a novel robust DP theory for continuous-time linear dynamical systems. Compared with traditional DP and adaptive optimal control, we take a completely different path to investigate DP methods from a viewpoint of nonlinear system theory [Kha02] and small-gain theory [Zam66, JTP94]. As a consequence, we will 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 by previously known results. The proposed robust DP framework is based on the dynamic property of differential matrix Riccati equation (DMRE). Recall from [Wil71, Kuč73] 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 [vdS17] for any arbitrarily large set of initial conditions within the region of attraction, which we will refer to as “semiglobal gain assignment”. This conclusion lays the foundation of our small-gain analysis on the continuous-time VI, which in turn leads to a sequence of convergence and robustness results. A comparison between different DP methods is given in table 1. We admit that one drawback of robust DP is that it requires the nominal value of the components in the algorithm, and hence is not model-free as in existing RL and adaptive optimal control 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 four classical problems arising from the field of adaptive optimal control. In the first application, we show that the continuous-time VI can be implemented with system matrices estimated iteratively from a time series. The estimation error is treated as an external disturbance, and the convergence of VI is proved via robust DP theory. In the second application, an improved version of the continuous-time adaptive dynamic programming (ADP) [BJ16b] is proposed by coupling the recursive least square (RLS) estimation of certain matrix inverse in the ADP learning process. Robust DP is used to tackle the presence of RLS error. Compared with existing results, the proposed ADP algorithm is more computationally efficient, as the estimate of the matrix inverse is updated together with the ADP learning. In the third application, we develop a continuous-time stochastic ADP theory for a class of ergodic control problems, that generalizes the main result of [BJJ16]. Different from the stochastic approximation [KY03] and Monte Carlo methods [SB98, Chapter 5] in traditional RL, a new method for convergence

analysis based on robust DP is proposed in the continuous-time setting, due to the complex nature of the continuous-time ergodic control problem. In our fourth and last application, we propose a novel decentralized VI algorithm for solving coupled AREs. The state-space based small-gain theory [JTP94] 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 algorithms for robust ADP [JJ13] and solving non-zero-sum differential games [SH69b, SH69a].

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

Notation: Throughout this paper, I_n denotes the identity matrix of dimension n . \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. $|\cdot|$ denotes the Euclidean norm for vectors, or the induced matrix norm for matrices. \mathcal{S}^n denotes the normed space of all n -by- n real symmetric matrices, equipped with the induced matrix norm. $\mathcal{S}_+^n = \{P \in \mathcal{S}^n : P > 0\}$. For a matrix $M \in \mathbb{R}^{n \times m}$, $\text{vec}(M) = [M_1^T, M_2^T, \dots, M_m^T]^T$, where $M_i \in \mathbb{R}^n$ is the i -th 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 $\text{vech}(M) = [M_{11}, M_{12}, \dots, M_{1n}, M_{22}, M_{23}, \dots, M_{(n-1)n}, M_{nn}]^T$, where $M_{ij} \in \mathbb{R}$ is the (i, j) -th element of matrix M . $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. \otimes and \oplus indicate the Kronecker product and Kronecker sum, respectively. Given a set Q , $\text{int}(Q)$ denotes the interior of Q . B_ε denotes an open ball centered at the origin with radius ε . A function $f : Q \rightarrow \mathbb{R}_+$, where $Q \subseteq \mathbb{R}^n$ and $0 \in Q$, is called positive definite, if $f(x) > 0$ for all $x \in Q \setminus \{0\}$, and $f(0) = 0$. For $f : \mathbb{R} \rightarrow \mathbb{R}_+$ and $g : \mathbb{R} \rightarrow \mathbb{R}_+$, denote $f(x) = o(g(x))$ if $\lim_{x \rightarrow 0} f(x)/g(x) = 0$.

2. PRELIMINARIES

2.1. System description. Consider the following linear time-invariant system:

$$(1) \quad \dot{x} = Ax + Bu,$$

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 eq. (1) as

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

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:

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

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

2.2. DMRE and continuous-time VI. Since eq. (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 [BJ16b]. Before introducing the VI algorithm, we define a real sequence $\{h_k\}_{k=0}^\infty$ satisfying

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

Algorithm 1 Continuous-time value iteration

Choose $P_0 = P_0^T > 0$. Let $\bar{\varepsilon} > 0$ be a small threshold. $k, q \leftarrow 0$.

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 $|P_{k+1/2}| > B_q$ or $P_{k+1/2} \not\geq 0$ **then**

$P_{k+1} \leftarrow P_0$. $q \leftarrow q + 1$.

else

$P_{k+1} \leftarrow P_{k+1/2}$

end if

$k \leftarrow k + 1$

end loop

In addition, denote $\{B_q\}_{q=0}^\infty$ as an increasing real sequence with $B_0 > 0$ and $\lim_{q \rightarrow \infty} B_q = \infty$.

The continuous-time VI is recalled from [BJ16b] 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 [BJ16b]. However, it still remains an open problem how robust algorithm 1 is to various types of disturbance. 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 [Wil71, Kuć73], for any $P(0) = P(0)^T \geq 0$, the solution to the following DMRE converges to P^* asymptotically as t goes to infinity:

$$(4) \quad \dot{P} = A^T P + P A - P B R^{-1} B^T P + Q.$$

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

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

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

$$(5) \quad \dot{\tilde{P}} = (A - B K^*)^T \tilde{P} + \tilde{P} (A - B K^*) - \tilde{P} B R^{-1} B^T \tilde{P}.$$

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

Lemma 3.1. \hat{P} is globally¹ exponentially stable at P^* , where \hat{P} is the solution of the following system:

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

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

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

Since $A - BK^*$ is Hurwitz, $((A - BK^*) \oplus (A - BK^*))^T$ is also Hurwitz [Bre78]. This completes the proof. \square

Remark 3.1. Note from eq. (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 .

Lemma 3.2. Consider a dynamical system defined on the inner product space $(\mathcal{S}^n, \langle \cdot, \cdot \rangle_F)$:

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

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

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

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

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

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 isomorphism². Then, one can rewrite eq. (8) as the following ODE:

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

where $p = \mathcal{M}(P)$, and $g = \mathcal{M} \circ G \circ \mathcal{M}^{-1}$. Denote the region of attraction of $0 \in \mathbb{R}^{n(n+1)/2}$ by $R'_A \subseteq \mathbb{R}^{n(n+1)/2}$. Since R_A is not empty, R'_A is also not empty. By converse Lyapunov theorem [Kha02, Theorem 4.17], we know there exists a smooth function $W(\cdot) : R'_A \rightarrow \mathbb{R}_+$, such that

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

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 eq. (9) since the solution to eq. (9) starting from $\mathcal{M}(P_0)$ also converges to the origin, by the norm preserving property of \mathcal{M} . 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 eq. (8).

¹Global 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.

²A bounded linear operator is called an isometric isomorphism if it is a norm preserving bijection which is continuous and has a continuous inverse [RS80, pp. 71].

Now, we define a function $V(\cdot) : R_A \rightarrow \mathbb{R}_+$, such that $V = W \circ \mathcal{M}$. By the definition of matrix calculus, $\partial_x V = \mathcal{M}^{-1} \circ \partial_x W \circ \mathcal{M}$. 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,

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

This concludes the proof. \square

Remark 3.2. lemma 3.2 extends the converse Lyapunov theorem for general nonlinear systems [Kha02, Theorem 4.17] to the space of real symmetric matrices. The converse statement of lemma 3.2, i.e., the Lyapunov theorem for the stability of general nonlinear systems over $(\mathcal{S}^n, \langle \cdot, \cdot \rangle_F)$, can also be derived in a similar way. Moreover, one can also generalize the converse Lyapunov theorem for exponentially stable systems [Kha02, Theorem 4.14]. We omit this direct extension to avoid duplication.

Proposition 3.3. *P is exponentially stable at P^* over \mathcal{S}^n .*

Proof. Note from lemma 3.1 and remark 3.1 that for any $\hat{P}(0) \in \mathcal{S}^n$, \hat{P} converges exponentially to P^* . Then, following lemma 3.2, remark 3.2, and eq. (7), there exists a smooth Lyapunov function $V : \mathcal{S}^n \rightarrow \mathbb{R}_+$ satisfying

$$\begin{aligned} C_1 |\hat{P} - P^*|^2 &\leq V(\hat{P} - P^*) \leq C_2 |\hat{P} - P^*|^2, \\ \dot{V}(\hat{P} - P^*) &\leq -C_3 |\hat{P} - P^*|^2, \quad |\partial_x V(\hat{P} - P^*)| < C_4 |\hat{P} - P^*|, \end{aligned}$$

for any \hat{P} on \mathcal{S}^n , where $C_i > 0$, $i = 1, 2, 3, 4$. Note that we use the induced norm here instead of the Frobenius norm, due to the equivalence of matrix norms.

Comparing the dynamics of \hat{P} and P , we see 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 eq. (5), we have

$$\dot{V}(\tilde{P}) \leq -C_3 |\tilde{P}|^2 + C_5 |\partial_x V(\tilde{P})| |\tilde{P}|^2 \leq -C_3 |\tilde{P}|^2 + C_4 C_5 |\tilde{P}|^3,$$

for any $\tilde{P} \in \mathcal{S}^n$ and constant $C_5 > 0$. From the above inequality, we know there exist $\varepsilon > 0$ and $C_6 > 0$, such that,

$$(10) \quad \dot{V}(\tilde{P}) \leq -C_6 |\tilde{P}|^2 \leq -\frac{C_6}{C_2} V(\tilde{P}), \quad \forall |\tilde{P}| < \varepsilon.$$

The proof is then completed using the Lyapunov theorem [Kha02, Theorem 4.10]. \square

Remark 3.3. Compared with [Wil71, Remark 21] and [Kuě73, Theorem 17], proposition 3.3 provides a stronger result, in the sense that it characterizes the convergence speed of $\tilde{P}(t)$ in a neighborhood of the origin. This is the foundation of our robustness analysis on DMRE and VI.

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

$$(11) \quad \dot{P}_\Delta = A^T P_\Delta + P_\Delta A - P_\Delta B R^{-1} B^T P_\Delta + Q + \Delta, \quad P_\Delta(0) = P_\Delta^T(0) \geq 0.$$

Remark 3.4. Δ can represent a large class of disturbances. In particular, we conduct robustness analysis on eq. (11) in theorem 3.4 below by considering three different forms of Δ , including *a*) a state-independent external signal (theorem 3.4, parts (i) and (ii)); *b*) the output of a nonlinear dynamical system (theorem 3.4, part (iii)); and *c*) a stochastic disturbance (theorem 3.4, 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 3.4. *Consider system eq. (11) with $Q > 0$. Denoting $\tilde{P}_\Delta = P_\Delta - P^*$, we have*

- (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 $\lim_{t \rightarrow \infty} \Delta(t) = 0$, then $\lim_{t \rightarrow \infty} P_\Delta(t) = P^*$. If in addition $\Delta \in L^2$, then $\tilde{P}_\Delta \in L^2$.*
- (iii) *There exists $\gamma > 0$, such that if the following system³*

$$(12) \quad \dot{M} = f(M, P_\Delta), \quad \Delta(t) = \Delta(P_\Delta, M),$$

where f and Δ are locally Lipschitz, $f(M^, P^*) = 0$, and $\Delta(P^*, M^*) = 0$, is zero-state detectable⁴ and admits an IOS Lyapunov function V_f satisfying*

$$(13) \quad \dot{V}_f(\tilde{M}) \leq -|\Delta|^2 + \gamma^2 |\tilde{P}_\Delta|^2, \quad \forall M \in B_{\varepsilon_0}(M^*), \quad \varepsilon_0 > 0,$$

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 $N > 0$, $\Delta_i : \mathcal{S}^n \rightarrow \mathcal{S}^n$, and the v_i are one-dimensional i.i.d. Gaussian white noises. Then, there exists $\gamma > 0$, such that if $\sum_i |\Delta_i|^2 < \gamma |\tilde{P}_\Delta|$ in a neighborhood of P^* , P_Δ is asymptotically stable at P^* in the mean square sense over \mathcal{S}^n .*

Proof. To prove part (i), we first introduce the following finite-horizon cost:

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

where $t < 0$ is an arbitrary time instant, and $Q(s) = Q + \Delta(s)$. Since $Q(s) \geq 0$ on $[t, 0]$, it is well known from the LQR theory [Lib12, Chapter 6.1] that $\inf_u \mathcal{J}_t(x(t); u, Q) = x^T(t)M(t)x(t)$, where $M(s) = M^T(s) > 0$, $s \in [t, 0]$, satisfies

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

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

³ 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.

⁴Here, with slight abuse of notation, we say eq. (12) is zero-state detectable if $\Delta \equiv 0$ and $P_\Delta \equiv P^*$ imply $M \equiv M^*$.

On the other hand, we have from the conditions on $Q + \Delta$ that there exists a constant matrix $\bar{Q} \in \mathcal{S}$, such that $0 \leq Q(s) < \bar{Q}$ for all s . Thus,

$$(14) \quad \begin{aligned} 0 \leq x^T(t)M(t)x(t) &\leq x^T(0)P_\Delta(0)x(0) + \int_t^0 (x^T Q(s)x + (\bar{u}^o)^T R \bar{u}^o) ds \\ &\leq x^T(0)P_\Delta(0)x(0) + \int_t^0 (x^T \bar{Q}x + (\bar{u}^o)^T R \bar{u}^o) ds, \end{aligned}$$

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

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

Then, we have from eq. (14) that $0 \leq M(t) < \bar{M}$ for all $t < 0$. Comparing the definitions of M and P_Δ , we know $M(t) = P_\Delta(-t)$. Thus, $0 \leq P_\Delta(t) < \bar{M}$ 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 $t > 0$. On the other hand, by [Son98, 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 $\lim_{t \rightarrow \infty} P_\Delta(t) = P^*$.

Moreover, choosing the same Lyapunov function in the proof of proposition 3.3, we know there exist positive constants C_1, C_2 , and ε_1 , such that

$$(15) \quad \dot{V}(\tilde{P}_\Delta) \leq -C_1|\tilde{P}_\Delta|^2 + C_2|\tilde{P}_\Delta||\Delta|, \quad \forall |\tilde{P}_\Delta| < \varepsilon_1,$$

where $\tilde{P}_\Delta = P_\Delta - P^*$. By completing the squares, we have from eq. (15) that eq. (11) admits a finite linear L^2 gain in a neighborhood of P^* . Thus, by H^∞ control theory, $\tilde{P}_\Delta \in L^2$ if $\Delta \in L^2$.

Now, we prove part (iii). Note from eq. (13) and eq. (15) that if

$$\gamma < \frac{C_1}{\sqrt{2}C_2},$$

then by defining $\bar{V}(P, M) = V(P) + \frac{C_2^2}{C_1}V_f(M)$,

$$\begin{aligned} \frac{d}{dt}\bar{V}(\tilde{P}_\Delta, \tilde{M}) &\leq -C_1|\tilde{P}_\Delta|^2 + C_2|\tilde{P}_\Delta||\Delta| - \frac{C_2^2}{C_1}|\Delta|^2 + \frac{C_2^2}{C_1}\gamma^2|\tilde{P}_\Delta|^2 \\ &= -\left(\frac{C_1}{2} - \frac{C_2^2}{C_1}\gamma^2\right)|\tilde{P}_\Delta|^2 - \frac{C_2^2}{2C_1}|\Delta|^2 \\ &\leq -C_3|\tilde{P}_\Delta|^2 - \frac{C_2^2}{2C_1}|\Delta|^2, \quad \forall |\tilde{P}_\Delta| < \varepsilon_1, |\tilde{M}| < \varepsilon_0, \end{aligned}$$

for some $C_3 > 0$. Since eq. (12) is zero-state detectable, we have from LaSalle's invariance principle [Kha02, 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 [Ste01] and eq. (10), it follows that

$$\mathcal{L}V(\tilde{P}_\Delta) \leq -C_1|\tilde{P}_\Delta|^2 + C_4 \sum_{i=1}^N |\Delta_i|^2, \quad \forall |\tilde{P}_\Delta| < \varepsilon,$$

for some positive constants ε and C_4 , where \mathcal{L} denotes the differential generator. Note that C_4 is bounded since $\partial_x^2 V$ is bounded on any compact sets, as V is smooth. Obviously, if

$$\sum_{i=1}^N |\Delta_i|^2 < \frac{C_1}{C_4} |\tilde{P}_\Delta|^2,$$

then

$$\mathcal{L}V(\tilde{P}_\Delta) \leq -C_5|\tilde{P}_\Delta|^2, \quad \forall |\tilde{P}_\Delta| < \varepsilon,$$

for some $C_5 > 0$. This concludes the proof. \square

proposition 3.3 and theorem 3.4 imply that the DMRE behaves very similar 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 these two theorems 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 eq. (2). Indeed, the following corollary shows that by choosing Q and R properly, we can guarantee the *semi-global* exponential stability of eq. (4) at P^* . By "semi-global", we mean that the domain of attraction is bounded but can be made as large as possible [Sas99].

Corollary 3.5. *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 eq. (4) starting at $P(0) \in \mathcal{S}_0$ converges exponentially to P^* .*

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 $\lim_{\lambda \rightarrow 0^+} P^* = 0$. 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 eq. (6), we have for any given $t > 0$ and $x(-t) \in \mathbb{R}^n$,

$$\begin{aligned} x^T(-t)P(t)x(-t) &= \inf_u \left\{ x^T(0)P(0)x(0) + \int_{-t}^0 (x^T Qx + u^T R u) ds \right\} \\ &\leq (x^*(0))^T P(0)x^*(0) + \int_{-t}^0 (x^*)^T (Q + (K^*)^T R K^*) x^* ds = x^T(-t)\hat{P}(t)x(-t), \end{aligned}$$

where x^* is the solution to system eq. (1) with $u = -K^*x^*$ and $x^*(-t) = x(-t)$. Moreover, by monotonicity [BJ16a, 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 $x^T P^* x$ exponentially for all x . Using $x^T P^* x \leq x^T P x \leq x^T \hat{P}x$, we know $x^T P x$ converges to $x^T P^* x$ exponentially. Noting that this is true for all x , P thus converges to P^* exponentially. This completes the proof. \square

Remark 3.5. It is easy to see from corollary 3.5 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 3.3 and the converse Lyapunov theorem [Kha02, 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 3.5 allows us to extend the result obtained in theorem 3.4 to any compact sets containing P^* in \mathcal{S}_+^n .

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

Corollary 3.6. *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 $\lambda > 0$, such that eq. (11) admits a finite linear L^2 gain from Δ to \tilde{P}_Δ 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 $R = o(\lambda)R_0$.*

Proof. Since Q_0 and R_0 are multiplied by different scaling factors, different from corollary 3.5, K^* depends on λ here. Hence, the first step of our proof is to characterize the influence of λ on the eigenvalues of $A - BK^*$.

Note that choosing $Q = \lambda Q_0$ and $R = o(\lambda)R_0$ is equivalent to choosing $Q = Q_0$ and $R = \delta_\lambda R_0$, where $\delta_\lambda = o(\lambda)/\lambda$, in the sense that these two choices lead to the same optimal controller. Denoting P_λ^* as the solution to eq. (3) with $Q = Q_0$ and $R = \delta_\lambda R_0$, we have

$$(16) \quad \begin{aligned} & (A - BK_\lambda^*)^T P_\lambda^* + P_\lambda^* (A - BK_\lambda^*) = \\ & - \left(Q_0 + \left(\sqrt{\delta_\lambda^{-1} P_\lambda^*} \right) B R_0^{-1} B^T \left(\sqrt{\delta_\lambda^{-1} P_\lambda^*} \right) \right), \end{aligned}$$

where $K_\lambda^* = \delta_\lambda^{-1} R_0^{-1} B^T P_\lambda^*$. Since B has full rank, we know from [KS72, (40)] that there exists $\bar{P} \in \mathcal{S}^n$, such that $\lim_{\lambda \rightarrow 0} \sqrt{\delta_\lambda^{-1} P_\lambda^*} = \bar{P}$. Thus, for any two positive constants C and ε , we can choose a small enough λ , such that $\left| \sqrt{\delta_\lambda^{-1} P_\lambda^*} - \bar{P} \right| < \varepsilon$ and

$$\sqrt{\delta_\lambda^{-1}} \left(Q_0 + \left(\sqrt{\delta_\lambda^{-1} P_\lambda^*} \right) B R_0^{-1} B^T \left(\sqrt{\delta_\lambda^{-1} P_\lambda^*} \right) \right) > C I_n.$$

This, together with the Lyapunov equation eq. (16), 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 implies that the eigenvalues of $A - BK_\lambda^*$ can be placed arbitrarily far to the left from the imaginary axis, by choosing a small enough λ .

Now, by the linear matrix inequality argument [GA94, Lemma 4.1], 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 v to ξ less than or equal to γ :

$$\dot{\xi} = ((A - BK_\lambda^*) \oplus (A - BK_\lambda^*))^T \xi + v.$$

Algorithm 2 Continuous-time robust VI

```

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  $P_{k+1/2} > 0$  and  $|P_{k+1/2} - P_k - h_k(\Delta_k + W_k)|/h_k < \bar{\varepsilon}$  then
    return  $P_k$  as an approximation to  $P^*$ 
  else if  $|P_{k+1/2}| > B_q$  or  $P_{k+1/2} \not\geq 0$  then
     $P_{k+1} \leftarrow P_0$ .  $q \leftarrow q + 1$ .
  else
     $P_{k+1} \leftarrow P_{k+1/2}$ 
  end if
   $k \leftarrow k + 1$ 
end loop

```

By following similar Lyapunov theorem arguments in the proof of proposition 3.3, we know that for any $\varepsilon > 0$, the L^2 gain of system eq. (11) can be made arbitrarily small on $\{P \in \mathcal{S}_+^n : P \in B_\varepsilon(P^*)\}$, by choosing a sufficiently small λ . This completes the proof. \square

Remark 3.6. 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 [JTP94, PW96, Isi99, LJH14]. 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 [KKK95], or a combination of the backstepping and small-gain approaches [LJH14].

The following corollary is a direct extension of theorem 3.4, parts (iii) and (iv), and corollary 3.6, and thus its proof is omitted.

Corollary 3.7. *Given $Q_0 = Q_0^T > 0$, $R_0 = R_0^T > 0$, and $\lambda > 0$, define $Q = \lambda Q_0$ and $R = o(\lambda)R_0$. Suppose B has full rank.*

- (i) *For any $\gamma > 0$, if system eq. (12) satisfies the conditions in theorem 3.4, part (iii), then there exist $\lambda > 0$, 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 3.4, 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 2) based on the theoretical results in Section 3.1. Note that different from 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 eq. (11).

Theorem 3.8. *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 Section 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 $\lim_{k \rightarrow \infty} \Delta_k = 0$ uniformly on any compact set in \mathcal{S}^n , then $\lim_{k \rightarrow \infty} P_k = P^*$.
- (iii) if $\Delta_k := \Delta(P_k, M_k)$ is the output to the following updating equation:

$$(17) \quad M_{k+1} = M_k + h_k f(M_k, P_k) + Z_k,$$

where $\{M_k\}_{k=0}^\infty$ is bounded in $B_{\varepsilon_0}(M^*)$ under a projection term Z_k , then there exists $\gamma > 0$, such that if the conditions in part (iii) of theorem 3.4 are satisfied, we have $\lim_{k \rightarrow \infty} (P_k, M_k) = (P^*, M^*)$ locally.

Proof. Before proving the part (i), we denote an operator $\mathcal{R} : \mathcal{S}^n \rightarrow \mathcal{S}^n$, such that

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

Suppose $P_0 \neq P^*$. By lemma 3.2, we know there exists a smooth Lyapunov function $\mathcal{V} : R_A \rightarrow \mathbb{R}_+$, where $R_A \subset \mathcal{S}^n$ is the region of attraction of P^* , such that

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

Note that \mathcal{V} defined here is different from the Lyapunov function used in the proof of proposition 3.3. 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 $|\zeta| < \varepsilon_\delta$,

$$(18) \quad \sup_{\{P: C_0 \leq \mathcal{V}(P) \leq C_1\}} \{\langle \partial_x \mathcal{V}(P), (\mathcal{R}(P) + \zeta) \rangle_F\} = -\delta,$$

for some $\delta > 0$.

By contradiction, suppose $\{P_k\}_{k=0}^\infty$ is unbounded. Then, there exists an up-crossing 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

$$(19) \quad \mathcal{V}(P_{k_j-1}) < C_2 \leq \mathcal{V}(P_m) < C_3 < \mathcal{V}(P_{k'_j}), \quad \forall k_j \leq m < k'_j.$$

Choose a sufficiently small $\varepsilon > 0$, such that for any $P \in \{P_{k_j}\}$, $B_\varepsilon(P) \subset \{P \in \mathcal{S}_+^n : \mathcal{V}(P) < C_1\}$. Suppose q is sufficiently large. Then, for any $j \in \mathbb{Z}_+$,

$$(20) \quad \begin{aligned} \varepsilon < |P_{L_\varepsilon(j)} - P_{k_j}| &= \left| \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i (\mathcal{R}(P_i) + \Delta_i + W_i) \right| \\ &\leq \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i (|\mathcal{R}(P_i)| + |\Delta_i| + |W_i|) \leq \varepsilon_C \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i, \end{aligned}$$

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

Then, by eq. (18) and the assumption on W_k , one has

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

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

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

Since $\lim_{j \rightarrow \infty} \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i W_i = 0$, 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 eq. (18) and eq. (20) it follows that

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

This implies 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 $k \leq i \leq k'$. 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

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

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:

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

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

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

where

$$\begin{aligned}
 H^k(t) &= \int_0^t (\mathcal{R}(P^k(s)) + \Delta^k(s)) ds, \quad Z^k(t) = \sum_{i=k}^{m(t+t_k)-1} Z_i, \\
 W^k(t) &= \sum_{i=k}^{m(t+t_k)-1} h_i W_i, \quad m(t) = \begin{cases} j, & 0 \leq t_j \leq t < t_{j+1}, \\ 0, & t < 0, \end{cases}
 \end{aligned}$$

$e^k(t)$ is due to replacing the second term on the right-hand side of the first equality in eq. (21) 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 [BJ16b, 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 [Sko56]. Following the procedure in the proof of [ABB02, 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 3.4, part (ii) and the Part 2 of the proof of [KY03, Theorem 5.2.1].

To prove part (iii), we note from the part (iii) of theorem 3.4 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 $\bar{V}(P, M) = \bar{V}(P - P^*, M - M^*)$, where the Lyapunov function \bar{V} is defined in the proof of theorem 3.4, we also have

$$\langle \partial_{x_1} \bar{V}(P, M), (\mathcal{R}(P) + \Delta + \zeta) \rangle_F + \langle \partial_{x_2} \bar{V}(P, M), f(M, P) \rangle_F < 0,$$

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 have (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 the part (iii) of theorem 3.4, this completes the proof. \square

Remark 3.7. The first two parts of theorem 3.8 focus on handling static uncertainties represented by either a bounded external disturbance input or a bounded function of P_k . The third part of theorem 3.8 deals with dynamic uncertainty, and hence is more suitable for developing decentralized VI algorithms.

The following corollary is a direct extension of corollary 3.7, part (i) and theorem 3.8, part (iii), and thus its proof is omitted:

Corollary 3.9. *Given $Q_0 = Q_0^T > 0$, $R_0 = R_0^T > 0$, and $\lambda > 0$, denote $Q = \lambda Q_0$ and $R = o(\lambda)R_0$. Suppose B has full rank. For any $\gamma > 0$, if the conditions in the part (iii) of theorem 3.4 are satisfied, then there exist $\lambda > 0$, such that $\lim_{k \rightarrow \infty} (P_k, M_k) = (P^*, M^*)$ locally, where M_k is defined in eq. (17).*

Remark 3.8. The boundedness of M_k can be relaxed, by extending the projection term in eq. (17) to the adaptive boundary case as in algorithm 2. The conclusions of theorem 3.8 and corollary 3.9 still hold, under minor changes of the proof.

The following corollary plays an important role in developing adaptive optimal control methods on the basis of the proposed robust VI framework.

Corollary 3.10. *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 $\lim_{k \rightarrow \infty} \Delta_k = 0$ uniformly on any compact set, σ_i are continuous, and v_k is a \mathcal{F}_k -adapted martingale difference with finite variance. Then, $\lim_{k \rightarrow \infty} P_k = P^*$ with probability one.*

Proof. We only need to show that P_k is bounded. Then, the convergence is proved by the part (ii) of theorem 3.8.

Again, by contradiction, suppose $\{P_k\}_{k=0}^{\infty}$ is unbounded. Following the analysis in the proof of theorem 3.8, 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 , and $L_\varepsilon(j)$ follow the same definitions in the proof of part (i) of theorem 3.8. Since $\lim_{k \rightarrow \infty} \Delta_k = 0$ uniformly on any compact set, $\sup_{i \in [k_j, L_\varepsilon(j)]} |\Delta_i(P_i)|$ can be made arbitrarily small, by choosing a large enough j . Then, there exists a sufficiently large \bar{j} , such that for all $j > \bar{j}$,

$$(22) \quad V(P_{L_\varepsilon(j)}) - V(P_{k_j}) \leq \langle \partial_x V(P_{k_j}), \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i \sigma_i(P_i) v_i \rangle_F - \delta \varepsilon / \varepsilon_C + O(\varepsilon^2).$$

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 eq. (19). Obviously, $\{M_k\}$ is a martingale with respect to $\{\mathcal{F}_k\}$, and $\mathbb{E}(|M_k|^2)$ is bounded, since P_i is bounded, $\sum_{k=0}^{\infty} h_k^2 < \infty$, and v_i has finite variance. By the martingale convergence theorem [Ste01, Theorem 2.6], M_k converges with probability one, and thus $\lim_{j \rightarrow \infty} \sum_{i=k_j}^{L_\varepsilon(j)-1} h_i \sigma_i(P_i) v_i = 0$. This, together with eq. (22), shows that P_k is bounded with probability one. \square

4. APPLICATIONS TO ADAPTIVE/STOCHASTIC/DECENTRALIZED OPTIMAL CONTROL

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

4.1. VI in the presence of modeling errors. Solving optimal control problems using algorithm 2 requires precise knowledge of system matrices. In practice, these system parameters may not be directly available, and are often estimated from measurement data subject to stochastic noise. In this subsection, we investigate the convergence of eq. (4) under estimated model parameters.

Suppose the system matrix A is not known *a priori*, and is approximated by a time series $\{\hat{A}(t)\}_{t \in \mathbb{R}_+}$, where $\hat{A}(t) = A + \sum_{i=1}^N \Delta_i v_i(t)$, $N > 0$, $\Delta_i \in \mathbb{R}^{n \times n}$ are constants, and v_i denote independent continuous-time Gaussian white noises.

Instead of eq. (4), let us consider the following equation:

$$\begin{aligned}
 \dot{P} &= \hat{A}^T P + P \hat{A} - P B R^{-1} B^T P + Q \\
 &= \mathcal{R}(P) + \sum_{i=1}^N (\Delta_i^T P + P \Delta_i) v_i \\
 (23) \quad &= \mathcal{R}(P) + \sum_{i=1}^N (\Delta_i^T \tilde{P} + \tilde{P} \Delta_i) v_i + \sum_{i=1}^N (\Delta_i^T P^* + P^* \Delta_i) v_i.
 \end{aligned}$$

Since Δ_i are constants, there exists a constant $\gamma > 0$, such that $\sum_{i=1}^N |\Delta_i^T P + P \Delta_i|^2 < \gamma |P|^2$. By theorem 3.4, part (iv), if γ is sufficiently small, then we can find a pair (Q, R) under which there exists a smooth Lyapunov function V satisfying

$$\begin{aligned}
 \mathcal{L}V(\tilde{P}) &\leq -C_1 |\tilde{P}|^2 + \gamma C_2 |\tilde{P}|^2 + \gamma C_2 |P^*|^2 \\
 &\leq -C_3 |\tilde{P}|^2 + \gamma C_2 |P^*|^2, \quad \forall P \in \{P \in \mathcal{S}_+^n : P \in B_\varepsilon(P^*)\},
 \end{aligned}$$

for some constants $\varepsilon > 0$ and $C_i > 0$, $i = 1, 2, 3$. Moreover, the noise-to-state stability gain [KD98] can be made arbitrarily small by choosing Q and R properly, if B has full rank.

The above inequality shows that the DMRE under noisy measurement of A is either locally or semi-globally (in \mathcal{S}_+^n) practically stable, with probability one. Indeed, due to the additive noise in eq. (23), $\lim_{t \rightarrow \infty} P(t)$ follows a steady state distribution.

To improve the convergence result, let's consider the following DMRE:

$$(24) \quad \dot{P} = \left(\frac{1}{t} \int_0^t \hat{A} ds \right)^T P + P \left(\frac{1}{t} \int_0^t \hat{A} ds \right) - P B R^{-1} B^T P + Q.$$

By definition,

$$\frac{1}{t} \int_0^t \hat{A} ds = A + \sum_{i=1}^N \Delta_i \frac{1}{t} \int_0^t v_i(s) ds = A + \sum_{i=1}^N \frac{1}{t} \Delta_i w_i(t),$$

where w_i are independent Brownian motions [Ste01, Chapter 3], and the last equality comes from the fact that v_i are independent Gaussian white noises. By the strong law of large number [Ste01, Appendix I], $\lim_{t \rightarrow \infty} \frac{1}{t} w_i(t) = 0$ with probability one.

Theorem 4.1. *Denote a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $\{\mathcal{F}_t\}_{t \geq 0}$. Suppose $(w_1(t), w_2(t), \dots, w_N(t))$ is \mathcal{F}_t -adapted. For any $P(0) \in \mathcal{S}_+^n$, we have $\lim_{t \rightarrow \infty} P(t) = P^*$ with probability one, where $P(t)$ is defined by eq. (24).*

Proof. By the definition of w_i , we know there exists $E \in \mathcal{F}$ with $\mathbb{P}(E) = 1$, such that for any $\omega \in E$, $\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^N \Delta_i w_i(\omega, t) = 0$. Now, fix a $\omega \in E$, and denote $\frac{1}{t} \sum_{i=1}^N \Delta_i w_i(\omega, t) := \Delta(t)$.

Define

$$\dot{\hat{P}} = (A - BK^*)^T \hat{P} + \hat{P}(A - BK^*) + K^* RK^* + Q + \Delta^T \hat{P} + \hat{P} \Delta, \quad \hat{P}(0) = P(0).$$

Since $\lim_{t \rightarrow \infty} \Delta(t) = 0$, we can easily show that \hat{P} is globally asymptotically stable at P^* on \mathcal{S}^n , by using the same Lyapunov function given in the proof of proposition 3.3. On the other hand, we have from eq. (24) that

$$\begin{aligned} 0 &\leq x^T(-t)P(t)x(-t) = \inf_u \left\{ x^T(0)P(0)x(0) + \int_{-t}^0 (x^T Q x + u^T R u) ds \right\} \\ &\leq (x^*(0))^T P(0) x^*(0) + \int_{-t}^0 (x^*)^T (Q + (K^*)^T R K^*) x^* ds = x^T(-t) \hat{P}(t) x(-t), \end{aligned}$$

where x is the solution to the following system:

$$\dot{x} = (A + \Delta)x + Bu.$$

Thus, $P(t)$ is bounded and stays in the region of attraction of P^* , for each given $\omega \in E$. The proof is then concluded following the similar analysis in the proof of theorem 3.4, part (ii). \square

In practice, instead of having $v_i(t)$, we usually have discrete-time white noise sequences $\{v_i(k)\}_{k=0}^{\infty}$ sampled from the continuous-time series, with constant variance σ_i^2 . In this case, eq. (23) and eq. (24) can be numerically approximated by

$$\begin{aligned} P_{k+1} &= P_k + h_k (\hat{A}_k^T P_k + P_k \hat{A}_k - P_k B R^{-1} B^T P_k + Q), \\ P_{k+1} &= P_k + h_k \left(\left(\frac{1}{k} \sum_{i=0}^k \hat{A}_i \right)^T P_k + P_k \left(\frac{1}{k} \sum_{i=0}^k \hat{A}_i \right) - P_k B R^{-1} B^T P_k + Q \right), \end{aligned}$$

respectively, where $P_0 = P(0)$, $\hat{A}_k = A + \sum_{i=1}^N \Delta_i v_i(k)$, and $h_k > 0$ is the step size. By the property of Gaussian white noise, we have $\lim_{h \rightarrow 0} \sum_{k=0}^{\lfloor t/h \rfloor} \sqrt{h} v_i(k) = \sigma_i w_i(t)$. Then the convergence of P_k can also be obtained.

Remark 4.1. Note that the system input matrix B can also be replaced by a time series $\hat{B}(t)$ in the above analysis.

4.2. VI-based ADP for linear continuous-time systems. ADP aims at solving the optimal control problem in real-time using the online input-state or input-output information. However, in traditional PI-based ADP methodologies, a matrix inverse is calculated in each learning iteration, which may induce a heavy computational burden in real world applications. Here, we solve this problem from the perspective of robust DP. This result is especially useful for high-order systems where solving matrix inverse online is not practical.

For all $x \in \mathbb{R}^n$ and $P \in \mathcal{S}^n$, taking the derivative along the solutions of system eq. (1), one has

$$(25) \quad \frac{d}{dt} \bar{x}^T \text{vech}(P) = \frac{d}{dt} (x^T P x) = (Ax + Bu)^T P x + x^T P (Ax + Bu) = \bar{z}^T \theta(P),$$

where $z = [x^T, u^T]^T$,

$$\theta(P) = \text{vech} \left(\begin{bmatrix} PA + A^T P & PB \\ B^T P & 0 \end{bmatrix} \right),$$

and for any $\xi \in \mathbb{R}^q$, $q \in \mathbb{Z}_+ \setminus \{0\}$,

$$\bar{\xi} = [\xi_1^2, 2\xi_1\xi_2, \dots, 2\xi_1\xi_q, \xi_2^2, 2\xi_2\xi_3, \dots, 2\xi_{q-1}\xi_q, \xi_q^2]^T.$$

Note that once $\theta(P)$ is obtained, we can define two linear transformations \mathcal{T}_A and \mathcal{T}_B , such that

$$A^T P + PA = \mathcal{T}_A(\theta(P)), \quad R^{-1}B^T P = \mathcal{T}_B(\theta(P)).$$

To provide an online implementation of algorithm 1, we need to solve $A^T P + PA$ and $R^{-1}B^T P$ from eq. (25) using online data only. First, define an arbitrary time sequence $0 \leq t_1 < t_2 < \dots < t_{l+1} < \infty$. Consider the following linear equation:

$$(26) \quad \psi_j^T(z)\theta(P) = \phi_j^T(x)\text{vech}(P), \quad \forall j \geq 0,$$

where

$$\phi_j(x) = \bar{x}(t_{j+1}) - \bar{x}(t_j), \quad \psi_j(z) = \int_{t_j}^{t_{j+1}} \bar{z} dt.$$

Now, by means of the RLS [Hay14, Chapter 10], one can define a sequence $\{\theta_k\}_{k=0}^l$ to approximate $\theta(P)$. To be specific, θ_k is updated by the following two equations:

$$(27) \quad \begin{aligned} \Sigma_k &= \Sigma_{k-1} - \frac{\Sigma_{k-1}\psi_k\psi_k^T\Sigma_{k-1}}{1 + \psi_k^T\Sigma_{k-1}\psi_k}, \\ \theta_k &= \theta_{k-1} + \Sigma_k\psi_k\phi_k^T\text{vech}(P) - \Sigma_k\psi_k\psi_k^T\theta_{k-1}, \end{aligned}$$

where $\theta_0 = 0$, and $\Sigma_0 = \lambda^{-1}I_q$ for $q = \frac{1}{2}((n+m)^2 + n+m)$ and some $\lambda > 0$.

Assumption 4.1. *There exist $l_0 > 0$ and $\alpha > 0$, such that*

$$(28) \quad \frac{1}{l} \sum_{j=1}^l \psi_j(z)\psi_j^T(z) > \alpha I$$

for all $l > l_0$.

If assumption 4.1 is satisfied, then we have from eq. (27) that

$$\begin{aligned} \theta_l &= \left(\sum_{j=1}^l \psi_j\psi_j^T + \lambda I_q \right)^{-1} \sum_{j=1}^l \psi_j\phi_j^T\text{vech}(P) \\ &= \left(\frac{1}{l} \sum_{j=1}^l \psi_j\psi_j^T + \frac{\lambda}{l} I_q \right)^{-1} \frac{1}{l} \sum_{j=1}^l \psi_j\phi_j^T\text{vech}(P), \end{aligned}$$

and thus

$$\lim_{l \rightarrow \infty} \theta_l = \theta(P).$$

By eq. (27) and using mathematical induction, we see θ_k is also linear in P . Hence one can find a matrix $M_k \in \mathbb{R}^{\left(\frac{n(n+1)}{2} + mn\right) \times \frac{n(n+1)}{2}}$, such that $\theta_k = M_k\text{vech}(P)$, with

Algorithm 3 Continuous-time VI-based ADP

Choose $P_0 = P_0^T \geq 0$, $\Sigma_0 = \lambda^{-1}I$, and $M_0 = 0$. $k, q \leftarrow 0$.
Apply a measurable locally essentially bounded input u to eq. (1).
loop
 $\Sigma_k \leftarrow \Sigma_{k-1} - \Sigma_{k-1}\psi_k\psi_k^T\Sigma_{k-1}/(1 + \psi_k^T\Sigma_{k-1}\psi_k)$
 $M_k \leftarrow M_{k-1} + \Sigma_k\psi_k(\phi_k^T - \psi_k^T M_{k-1})$
 $\hat{\theta}_k \leftarrow M_k \text{vech}(P_k)$
 $P_{k+1/2} \leftarrow P_k + h_k(\mathcal{T}_A(\hat{\theta}_k) - \mathcal{T}_B^T(\hat{\theta}_k)R\mathcal{T}_B(\hat{\theta}_k) + Q)$
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 $|P_{k+1/2}| > B_q$ or $P_{k+1/2} \not\geq 0$ **then**
 $P_{k+1} \leftarrow P_0$. $q \leftarrow q + 1$.
else
 $P_{k+1} \leftarrow P_{k+1/2}$
end if
 $k \leftarrow k + 1$
end loop

$M_0 = 0$, for all $k = 0, 1, \dots, l$. Moreover, since eq. (25) is true for any $P \in \mathcal{S}^n$, by replacing θ_k in eq. (27) with $M_k \text{vech}(P)$, we must have

$$(29) \quad M_k = M_{k-1} + \Sigma_k\psi_k\phi_k^T - \Sigma_k\psi_k\psi_k^T M_{k-1}.$$

By the convergence of θ_l , we have

$$\lim_{l \rightarrow \infty} M_l = \lim_{k \rightarrow \infty} M_k = M,$$

where M satisfies $\theta(P) = M \text{vech}(P)$ for all $P \in \mathcal{S}^n$.

Based on the above analysis, the VI-based ADP algorithm for linear continuous-time systems is given in algorithm 3.

Theorem 4.2. *Under assumption 4.1, we have $\lim_{k \rightarrow \infty} P_k = P^*$ and $\lim_{k \rightarrow \infty} K_k = K^*$, where $\{P_k\}_{k=0}^\infty$ is obtained from algorithm 3, and $K_k = \mathcal{T}_B(\hat{\theta}_k)$.*

Proof. Noting that

$$\theta(P_k) = M \text{vech}(P_k) = \hat{\theta}_k + (M - M_k) \text{vech}(P_k),$$

we have

$$\begin{aligned} \mathcal{T}_A(\hat{\theta}_k) &= A^T P_k + P_k A + \Delta_{1,k}(P_k), \\ \mathcal{T}_B(\hat{\theta}_k) &= R^{-1} B^T P_k + \Delta_{2,k}(P_k), \end{aligned}$$

for some linear functions $\Delta_{1,k}$ and $\Delta_{2,k}$. Thus, algorithm 3 is essentially a special case of algorithm 2 with

$$\Delta_k = \Delta_{1,k}(P_k) + \Delta_{2,k}^T(P_k)R\Delta_{2,k}(P_k) + \Delta_{2,k}^T(P_k)B^T P_k + P_k B \Delta_{2,k}(P_k).$$

Since $\lim_{k \rightarrow \infty} M_k = M$ under assumption 4.1, both $\Delta_{1,k}$ and $\Delta_{2,k}$ converge to zero over any compact set. Then, the convergence of P_k to P^* is proved by theorem 3.8, part (ii). Following the definition, we then easily have $\lim_{k \rightarrow \infty} K_k = K^*$. This completes the proof. \square

Remark 4.2. Since the RLS scheme is also robust to stochastic noise [Hay14, Chapter 10], it is possible to extend algorithm 3 to the stochastic optimal control framework.

4.3. Stochastic ADP for ergodic control problems. In this subsection, we develop an ADP algorithm to solve the ergodic control problem [Bor06] for linear stochastic systems with additive noise.

Consider the following system:

$$(30) \quad dx = Axdt + Budt + \sum_{i=1}^{q_x} \sigma_i dw_{x,i},$$

$$(31) \quad du = -K_0 dx + \sum_{i=1}^{q_u} \sigma_{u,i} dw_{u,i},$$

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

Remark 4.3. $\sum_{i=1}^{q_x} \sigma_i dw_{x,i}$ in eq. (30) represents the additive noise in system eq. (30). $\sum_{i=1}^{q_u} \sigma_{u,i} dw_{u,i}$ in eq. (31) serves as an exploration noise, which has been widely used in adaptive control literature to guarantee the persistent excitation condition (PE) [Tao03, 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 eq. (31) here, as in this case system eq. (30) is purely driven by Brownian motions, and several standard results from SDE 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 eq. (30)-eq. (31).*

A discrete time version of assumption 4.2 for MDPs has been widely used in approximate DP and RL literature [Tsi94, TVR97]. See [Won67, Hau71] for conditions under which assumption 4.2 holds.

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

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

where $Q = Q^T > 0$ and $R = R^T > 0$. It can be shown [Bor06] that $\inf_u \mathcal{J}(u) = \sum_{i=1}^{q_x} \sigma_i^T P^* \sigma_i$, with P^* and the optimal controller sharing the same definitions as the ones in Section 2.1 for deterministic systems.

Now, we derive an online ADP algorithm to solve the above ergodic control problem. Similar to eq. (25), for all $x \in \mathbb{R}^n$ and $P \in \mathcal{S}^n$, by Itô's lemma [Ste01, Theorem 8.3], we have along the trajectories of eq. (30) that

$$(32) \quad \begin{aligned} d(x^T P x) &= 2x^T P (Ax + Bu) dt + \sum_{i=1}^{q_x} \sigma_i^T P \sigma_i dt + 2x^T P \sum_{i=1}^{q_x} \sigma_i dw_{x,i} \\ &= \psi^T(x, u) \theta(P) dt + 2x^T P \sum_{i=1}^{q_x} \sigma_i dw_{x,i}, \end{aligned}$$

where $\psi(x, u) = [\bar{x}^T, x^T \otimes u^T, 1]^T$, and

$$\theta(P) = \begin{bmatrix} \text{vech}(PA + A^T P) \\ \text{vec}(B^T P) \\ \sum_{i=1}^{q_x} \sigma_i^T P \sigma_i \end{bmatrix}.$$

Then, multiplying ψ on both sides of eq. (32), we have on any finite time interval $[0, T]$ that

$$(33) \quad \frac{1}{T} \int_0^T \psi \psi^T dt \theta(P) = \frac{1}{T} \int_0^T \psi d(x^T P x) - \frac{2}{T} \int_0^T \psi x^T P \sum_{i=1}^{q_x} \sigma_i dw_{x,i}.$$

Once $\theta(P)$ is obtained from the equation above, we can use the linear transformations defined similarly as the ones in Section 4.2 to find $A^T P + PA$ and $R^{-1} B^T P$:

$$A^T P + PA = \mathcal{T}_A(\theta(P)), \quad R^{-1} B^T P = \mathcal{T}_B(\theta(P)).$$

In order to solve eq. (33), we impose the following assumption:

Assumption 4.3. μ satisfies

$$(34) \quad \int_{\mathbb{R}^n \times \mathbb{R}^m} \psi \psi^T d\mu > 0.$$

Note that system eq. (30)-eq. (31) is a multidimensional Ornstein-Uhlenbeck process. Then, its stationary probability measure μ is also Gaussian, and thus (x, u) has finite r -th moment for any $r \in \mathbb{N}_+$. assumption 4.3 is similar to the PE condition widely used in adaptive control literature (see remark 4.3 for details).

By a direct extension of Birkhoff's ergodic theorem [ABG12, Theorem 1.5.18] and the Itô's isometry [Ste01, Theorem 6.1], we know³

$$(35) \quad \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{P}} \left[\left\| \frac{1}{T} \int_0^T \psi \psi^T dt - \int_{\mathbb{R}^n \times \mathbb{R}^m} \psi \psi^T d\mu \right\|_2^2 \right] = 0,$$

$$(36) \quad \begin{aligned} & \lim_{T \rightarrow \infty} \mathbb{E}^{\mathbb{P}} \left[\left\| \frac{1}{T} \int_0^T \psi d(x^T P x) - \frac{1}{T} \int_0^T \psi \psi^T dt \theta(P) \right\|_2^2 \right] \\ & = \lim_{T \rightarrow \infty} \frac{4}{T^2} \mathbb{E}^{\mathbb{P}} \left[\left\| \int_0^T \psi x^T P \sum_{i=1}^{q_x} \sigma_i dw_{x,i} \right\|_2^2 \right] = 0. \end{aligned}$$

Choosing a monotone increasing sequence $\{t_k\}_{k=0}^{\infty}$ with $t_0 > 0$ and $\lim_{k \rightarrow \infty} t_k = \infty$, we denote

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

Note from eq. (34) and eq. (35) that t_0 is a stopping time, and $t_0 < \infty$ with probability one, such that $\int_0^{t_k} \psi \psi^T dt$ is invertible for all $k \geq 0$.

For simplicity, denote $\hat{\theta}_k = \hat{\theta}(P, t_k)$. The VI-based ADP algorithm for the ergodic control problem is given in algorithm 4.

³ $\|\cdot\|_2$ 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 4 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 eq. (31).
loop
 $\hat{\theta}_k = \left(\int_0^{t_k} \psi \psi^T dt \right)^{-1} \int_0^{t_k} \psi d(x^T P_k x)$.
 $P_{k+1/2} \leftarrow P_k + h_k (\mathcal{T}_A(\hat{\theta}_k) - \mathcal{T}_B^T(\hat{\theta}_k) R \mathcal{T}_B(\hat{\theta}_k) + Q)$
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 $|P_{k+1/2}| > B_q$ or $P_{k+1/2} \not\geq 0$ **then**
 $P_{k+1} \leftarrow P_0$. $q \leftarrow q + 1$.
else
 $P_{k+1} \leftarrow P_{k+1/2}$
end if
 $k \leftarrow k + 1$
end loop

Theorem 4.3. Under assumption 4.2 and assumption 4.3, we have $\lim_{k \rightarrow \infty} P_k = P^*$ and $\lim_{k \rightarrow \infty} K_k = K^*$ with probability one, where $\{P_k\}_{k=0}^\infty$ is obtained from algorithm 4, and $K_k = \mathcal{T}_B(\hat{\theta}_k)$.

Proof. First denote

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

Then, by eq. (35) and eq. (36), we have

$$\lim_{k \rightarrow \infty} \mathbb{E}^{\mathbb{P}} [\|\Delta_k(P)\|_2^2] = 0.$$

Since the above formulation is true for any real symmetric P , we know for any $P \in \mathcal{S}^n$, $\Delta_k(P)$ is a martingale (element wise), and converges to 0 as k goes to infinite, in the mean square sense.

By Burkholder-Davis-Gundy inequality [BDG72, 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 $[\cdot]$ denotes the quadratic variation [Ste01, Section 8.6]. By Birkhoff's ergodic theorem and the fact that (x, u) has finite r -th moment for any $r \in \mathbb{N}_+$, $\lim_{k \rightarrow \infty} [\Delta^{i,j}(P)]_k = 0$ with probability one. Hence $\lim_{k \rightarrow \infty} \mathbb{E}^{\mathbb{P}} \left[[\Delta^{i,j}(P)]_k^2 \right] = 0$. This implies that the variance of $\Delta_k^T \Delta_k$ is bounded.

Now, the updating equation in algorithm 4 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 3.10. \square

Remark 4.4. It is possible to extend the results in this section to systems with both multiplicative and additive noises:

$$dx = Axd t + Bud t + \sum_{i=1}^{q_x} \sigma_i dw_{x,i} + \sum_{i=1}^{q_1} F_i x dw_{1,i} + \sum_{i=1}^{q_2} G_i u dw_{2,i},$$

where F_i and G_i are constant matrices. In this case, the ARE is given as

$$A^T P^* + P^* A + \sum_{i=1}^{q_1} F_i^T P^* F_i - P^* B \left(R + \sum_{i=1}^{q_2} G_i^T P^* G_i \right)^{-1} B^T P^* + Q = 0.$$

The convergence of VI in this case is guaranteed using results in [BJ16b, Theorem 3.3] and [ARCMZ01]. The robust VI and ADP algorithms similar to algorithm 2 and algorithm 4 can be derived following the analysis given before.

4.4. Decentralized VI. In previous sections, we have studied different types of optimal control problems for continuous-time linear systems. A common feature in 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 solve a linear optimal control problem (see Section 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 without communicating with the other one. However, assume now 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 relationship $\Delta_i(\cdot)$, and for security reason the two agents cannot exchange their system information (A_i, B_i, Q_i, R_i) , $i = 1, 2$, to each other, 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:

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

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

Assumption 4.4. *There exist four polynomials $\gamma_{i,j} \in \mathcal{K}$, $i, j = 1, 2$, such that⁴*

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

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 $\tilde{P}_2 = P_2 - P_2^*$.

⁴A function $\gamma : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is of class \mathcal{K} , if it is continuous, strictly increasing, and $\gamma(0) = 0$.

Remark 4.5. assumption 4.4 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 [JJ17, 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.

Theorem 4.4. *Under assumption 4.4, there exist $\varepsilon > 0$ and small enough $|\gamma_{i,j}|$, $i, j = 1, 2$, such that given $(P_1(0), P_2(0))$ in a ε -neighborhood of (P_1^*, P_2^*) , we have $\lim_{t \rightarrow \infty} P_1(t) = P_1^*$ and $\lim_{t \rightarrow \infty} P_2(t) = P_2^*$, where*

$$(37) \quad \dot{P}_1 = A_1^T P_1 + P_1 A_1 - P_1 B_1 R_1^{-1} B_1^T P_1 + Q_1 + \Delta_1(P_1, P_2),$$

$$(38) \quad \dot{P}_2 = A_2^T P_2 + P_2 A_2 - P_2 B_2 R_2^{-1} B_2^T P_2 + Q_2 + \Delta_2(P_2, P_1).$$

Moreover, if B_i has full rank, the convergence result holds for any $\gamma_{i,j}$ by picking Q_i and R_i properly.

Proof. Following the derivation of eq. (15), 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 decrease to 0 at least as fast as the first term. Hence, eq. (37) and eq. (38) are asymptotically stable at (P_1^*, P_2^*) , as long as the gain of $\gamma_{i,j}$ is small enough.

Moreover, if B_1 has full rank, we know from corollary 3.6 that eq. (37) can have an arbitrarily small linear L^2 gain from $\tilde{\Delta}_1$ to \tilde{P}_1 , i.e., C_2/C_1 can be made sufficiently small, on any compact sets, by choosing Q_1 and R_1 properly. If $|\gamma_{2,2}|$ is sufficiently small, then for any $\varepsilon > 0$, we can find Q_1 and R_1 , such that

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

for some $C_5 > 0$. This completes the proof. \square

Based on theorem 4.4, we develop a coupled VI algorithm in algorithm 5. The convergence of algorithm 5 is given in the following theorem.

Algorithm 5 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 approximations to P_i^*

end if

$k \leftarrow k + 1$

end loop

Theorem 4.5. *Under assumption 4.4, suppose 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 $R_i = o(\lambda_i) R_{i,0}$, we have $\lim_{k \rightarrow \infty} P_{i,k} = P_i^*$, where $\{P_{i,k}\}_{k=0}^\infty$ is obtained from algorithm 5 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 small, we know from part (i) of corollary 3.7 that the couple system eq. (37) and eq. (38) 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 also from corollary 3.6 that ε can be made arbitrarily large.

Now, choosing $\sup_k \{h_{i,k}\}$ sufficiently small, we easily have from part (i) of theorem 3.8 that $\{P_{i,k}\}$ stays in $\mathcal{S}_+^{n_i} \cap B_\varepsilon(P_i^*)$. Then, the proof is completed by part (iii) of theorem 3.8. \square

Remark 4.6. The results presented in this section can be extended in different directions, such as for large-scale networks with more than two nodes and decentralized VI under stochastic disturbance.

5. ILLUSTRATIVE PRACTICAL EXAMPLES

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

5.1. Mean-variance portfolio optimization. In this example, we study the mean-variance portfolio optimization problem [ZL00] using non-zero-sum differential game theory and the robust VI results obtained in Sections Section 4.1 and Section 4.4.

Consider the price process of $N + 1$ assets (or securities) traded continuously in a market [ZL00]:

$$\begin{aligned} \frac{dS_0}{S_0} &= r dt, \\ \frac{dS_i}{S_i} &= b_i dt + \sum_{j=1}^{n_i} \sigma_{ij} dw_j, \quad i = 1, 2, \dots, N, \end{aligned}$$

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 i -th stock. An investor's total wealth at time t , when holding $h_i(t)$ shares of

the i -th asset, is given as

$$x(t) = \sum_{i=0}^N h_i(t) S_i(t).$$

Then,

$$dx = \left(rx + \sum_i (b_i - r) u_i \right) dt + \sum_{i,j} \sigma_{ij} u_i dw_j,$$

where $u_i := h_i S_i$ denotes the total market value of the investor's wealth in the i -th bond/stock. The design objective here is to find u to a) maximize the average return; and b) minimize the volatility of x .

Inspired by [ZL00], instead of solving the above portfolio optimization problem directly, we consider an auxiliary multi-player non-zero-sum differential game composed with the following cost

$$(39) \quad J_i(u) = \mathbb{E} \left[\int_0^\infty \left(Q_i \bar{x}^2 + \sum_{j=1}^N R_{ij} \bar{u}_i \bar{u}_j \right) dt \right], \quad i = 1, \dots, N,$$

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 eq. (39) is related to the variance of \bar{x} (and hence x) at the steady state, and the second term guarantee that the shares for the i -th 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 [BJ16b, 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 5, 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 i -th asset at time t is chosen as $K_i^*(\gamma - \bar{x}(t))$. Totally 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 0–15%. Suppose the real values of these rates are unknown, and are estimated online using techniques developed in Section 4.1. After 1000 iterations, all P_i^* converge to their optimal values. The prices of the portfolio is shown in fig. 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 learning for kinematic models. In this example, we use the ADP method proposed in Section 4.2 to develop an online learning mechanism for a class

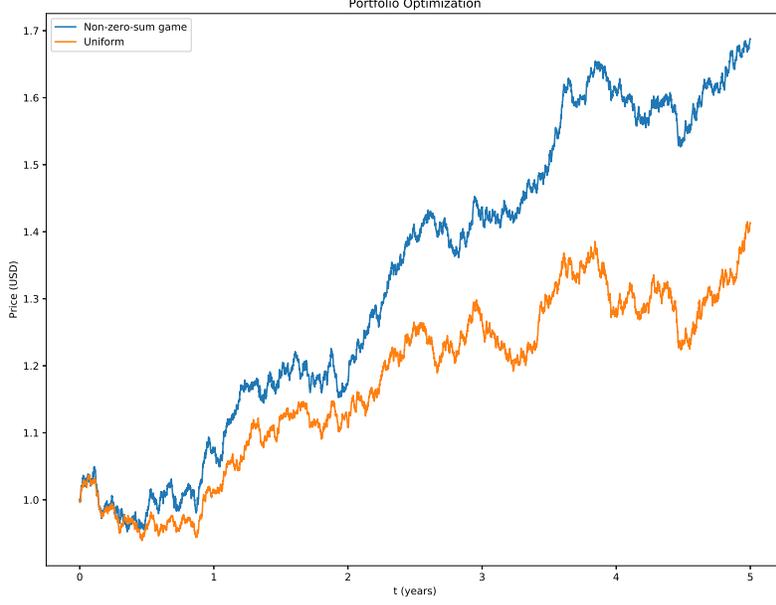


FIGURE 1. Example 5.1: Portfolio optimization solved from robust VI.

of kinematic models described as follows:

$$\begin{aligned}\dot{p} &= v, \\ m\dot{v} &= f - bv, \\ \tau\dot{f} &= u - f + w,\end{aligned}$$

where p , v , f denote the relative position to the origin, the velocity, and the actuator force, respectively; u is the control input; m , b , and τ represent the mass, the viscosity constant, and the time constant, respectively; and w is an exploration noise used to facilitate the ADP learning. Note that the above system can represent a large class of practical systems, including human motor system, autonomous vehicle model, power system, to name a few.

In practice, the state information collected from online data is usually corrupted by some observation noises. As a result, instead of (p, v, f) , we assume only $(\hat{p}, \hat{v}, \hat{f})$ is observed and used in the feedback control design and ADP learning:

$$\begin{aligned}\hat{p} &= p + \sigma_p w_p, \\ \hat{v} &= v + \sigma_v w_v, \\ \hat{f} &= f + \sigma_f w_f,\end{aligned}$$

where σ_p , σ_v , and σ_f denote the noise magnitude; and w_p , w_v , and w_f are independent random variable and follow standard Gaussian distribution.

TABLE 2. Parameters of the Kinematic Model.

Parameters	Description	Value
m	Mass	1kg
b	Viscosity constant	1N·s/m
τ	Time constant	0.1s
σ_p	Noise magnitude	0.01
σ_v	Noise magnitude	0.02
σ_f	Noise magnitude	0.1

The values of model parameters used in simulation are provided in table 2. algorithm 3 is applied online, and the control policy is updated in real time after every 0.02s. The weighting matrices in the cost are chosen as $Q = I_3$ and $R = I_1$. The initial controller $u \equiv 0$, i.e., only the exploration noise is injected into the system at the beginning. The elements in P_k are plotted in fig. 2 for each k . For comparison purpose, both the optimal solution P^* and the near optimal solution \hat{P}^* learned through ADP are given below:

$$P^* = \begin{bmatrix} 7.4044 & 1.4311 & 0.1000 \\ 1.4311 & 0.3801 & 0.0248 \\ 0.1000 & 0.0248 & 0.0431 \end{bmatrix}, \quad \hat{P}^* = \begin{bmatrix} 7.4560 & 1.5184 & 0.1084 \\ 1.5184 & 0.5117 & 0.0216 \\ 0.1084 & 0.0216 & 0.0463 \end{bmatrix}.$$

Obviously, P^* and \hat{P}^* are close to each other. The system trajectories and input are given in fig. 3. Note that the system achieves the asymptotical stability in mean square sense.

5.3. ADP for time-series variance minimization. In this example, we use the ADP method developed in Section 4.3 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 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:

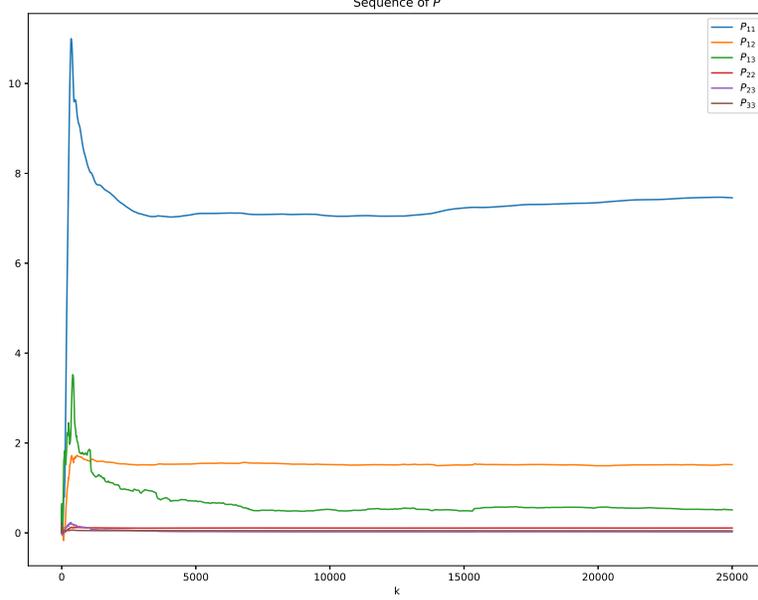
$$\ddot{S} = \alpha_3 \dot{S} + \alpha_2 \dot{S} + \alpha_1 S + \sigma_0 v_0,$$

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 .

By rewriting the above differential equation in state space form, we have

$$\begin{aligned} dx_1 &= x_2 dt + \sigma_1 dw_1 - \sigma_2 w_2 dt, \\ dx_2 &= x_3 dt + \sigma_2 dw_2 - \sigma_3 w_3 dt, \\ dx_3 &= \alpha_3 x_3 dt + \alpha_2 x_2 dt + \alpha_1 x_1 dt + \sigma_3 dw_3 + u + \sigma_0 dw_0, \end{aligned}$$

where $x_1 = S + \sigma_1 w_1$, $x_2 = \dot{S} + \sigma_2 w_2$, $x_3 = \ddot{S} + \sigma_3 w_3$; $w_0 = \dot{v}_0$; w_i , $i = 0, 1, 2, 3$, are Brownian motions representing the observation noises; σ_i , $i = 1, 2, 3$, are unknown noise magnitudes; and u is the control input. 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 4 is used to develop an ergodic controller. Notice that by the law of large

FIGURE 2. Example 5.2: elements of P_k .

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 4.

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 purpose, the weighting matrices in the cost are chosen as $Q = 0.1I_3$ and $R = 0.01I_1$. P_k is updated in real time after every 1s. The elements in P_k are given in fig. 4. 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 are given in fig. 5. Note that the controller derived from algorithm 4 significantly reduces the variance of the output signal.

6. SUMMARY AND FUTURE WORK

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

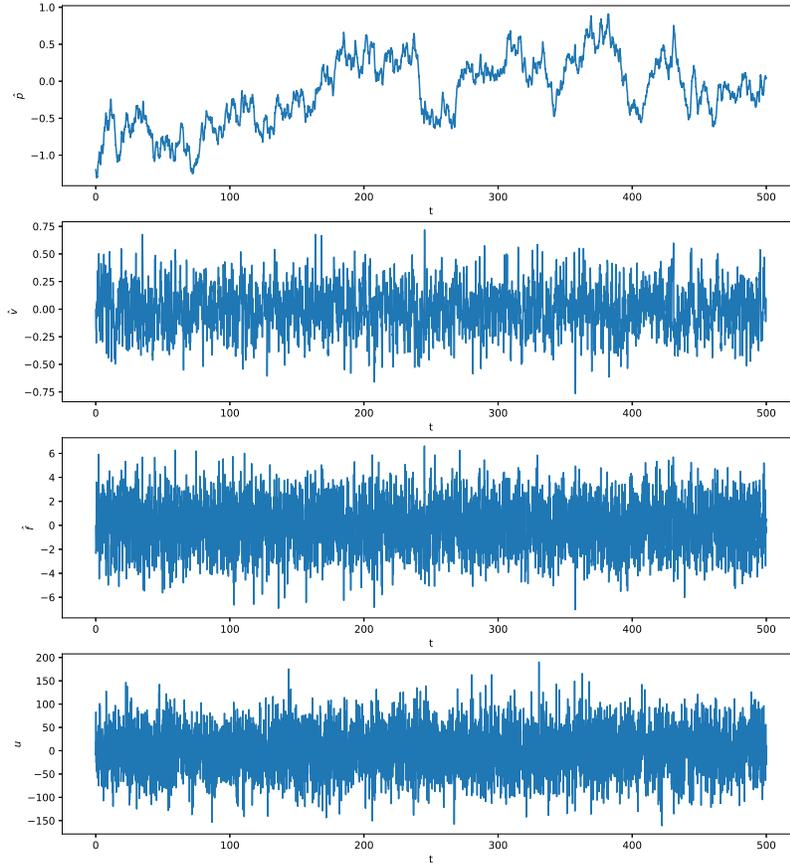


FIGURE 3. Example 5.2: the trajectories of \hat{p} , \hat{v} , \hat{f} , and u .

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 limitations, we only list 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. Last but not least, we point out several additional

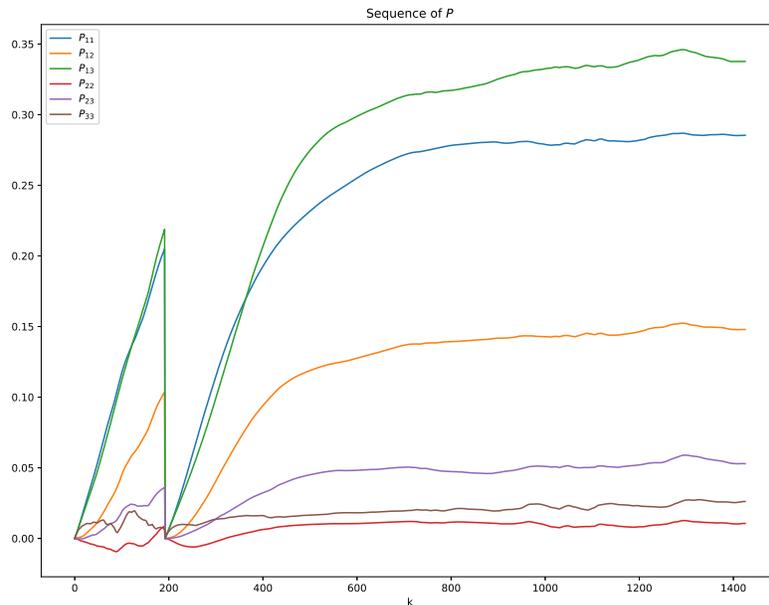


FIGURE 4. Example 5.3: elements of P_k .

research topics that deserve further investigations in the future on the basis of robust DP:

- *Robust PI.* This paper mainly focuses on developing the robust VI. Due to the popularity of the PI algorithm in real-world decision making applications, it is important to develop similar robustness analysis results for the PI.
- *Multi-level ADP learning.* The convergence of ADP algorithms relies heavily on the well-posedness of the cost functional. However, it may not be easy to identify such a “qualified” cost in practice. One way of solving this problem is to update the cost functional at the same time when the ADP learning is performed. The convergence of this multi-level learning algorithm can be analyzed using our robust DP framework.
- *Neural network-based ADP methods.* Robust DP can also play an important role in analyzing the convergence of nonlinear ADP methods with neural network approximation. The error induced from neural network approximation can be regarded as an external input to the robust DP algorithm. Then the gain analysis can be conducted to quantify the influence of such approximation error.
- *Robust ADP learning under unknown disturbance.* Robust ADP aims at developing a robust adaptive optimal controller for an interconnected system subject to dynamic uncertainty. A potential drawback of previous robust ADP algorithms is that the disturbance input must be accessible during

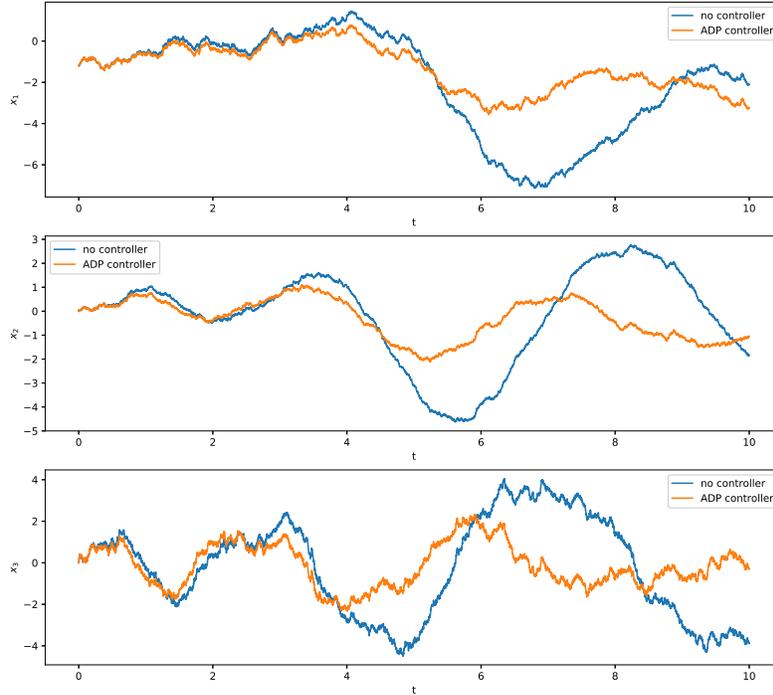


FIGURE 5. Example 5.3: the trajectories of x_i , $i = 1, 2, 3$.

the learning process. This restrictive assumption can be removed with the help of the proposed robust DP methodology.

- *DP and ADP methods for delayed systems.* The time delay can be handled as a special type of dynamic uncertainty with the unity gain. Since robust DP provides a new way of conducting the convergence analysis from a nonlinear small-gain perspective, it can play a vital role in handling the adaptive optimal control design with delayed input and state information.
- *Parallel and decentralized ADP methods.* Section 4.4 has developed a decentralized VI algorithm that is shown especially useful in solving large-scale DP problems and differential games. Our future work will be directed at developing the model-free counterpart of algorithm 5, and extending this result to more general scenarios.

ACKNOWLEDGEMENTS

The work of Z.P. Jiang has been supported partially by the National Science Foundation under Grants ECCS-1230040 and ECCS-1501044.

REFERENCES

- [ABB02] J. Abounadi, Dimitri P. Bertsekas, and Vivek S. Borkar. Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms. *SIAM Journal on Control and Optimization*, 41(1):1–22, 2002.
- [ABG12] Aristotle Arapostathis, Vivek S. Borkar, and Mrinal K. Ghosh. *Ergodic Control of Diffusion Processes*. Cambridge University Press, New York, NY, 2012.
- [ARCMZ01] M. Ait Rami, Xi Chen, J. B. Moore, and Xun Yu Zhou. Solvability and asymptotic behavior of generalized Riccati equations arising in indefinite stochastic lq controls. *IEEE Transactions on Automatic Control*, 46(3):428–440, Mar 2001.
- [BDG72] 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*, volume 2, pages 223–240, Berkeley, CA, 1972.
- [Bea95] Randal W. Beard. *Improving the closed-loop performance of nonlinear systems*. PhD thesis, Rensselaer Polytechnic Institute, 1995.
- [Bel52] Richard Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716–719, 08 1952.
- [Bel57] Richard E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.
- [Ber05] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, Belmont, MA, 3rd edition, 2005.
- [Ber07] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 3rd edition, 2007.
- [Ber13] Dimitri P. Bertsekas. *Abstract Dynamic Programming*. Athena Scientific, Belmont, MA, 2013.
- [Ber17] Dimitri P. Bertsekas. Value and policy iterations in optimal control and adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):500–509, March 2017.
- [BJ16a] Tao Bian and Z. P. Jiang. Value iteration, adaptive dynamic programming, and optimal control of nonlinear systems. In *Proceedings of the 55th IEEE Conference on Decision and Control (CDC)*, pages 3375–3380, Las Vegas, USA, 2016.
- [BJ16b] Tao Bian and Z. P. Jiang. Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. *Automatica*, 71:348–360, 2016.
- [BJJ14] Tao Bian, Yu Jiang, and Z. P. Jiang. Adaptive dynamic programming and optimal control of nonlinear nonaffine systems. *Automatica*, 50(10):2624–2632, 2014.
- [BJJ16] Tao Bian, Yu Jiang, and Z. P. Jiang. Adaptive dynamic programming for stochastic systems with state and control dependent noise. *IEEE Transactions on Automatic Control*, 61(12):4170–4175, Dec 2016.
- [Bor06] Vivek S. Borkar. Ergodic control of diffusion processes. In *Proceedings of the International Congress of Mathematicians*, pages 1299–1309, 2006.
- [Bre78] John W. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, 25(9):772–781, 1978.
- [BT96] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.
- [BTS17] Andrew G. Barto, Philip S. Tomas, and Richard S. Sutton. Some recent applications of reinforcement learning. In *Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems*, 2017.
- [GA94] Pascal Gahinet and Pierre Apkarian. A linear matrix inequality approach to H^∞ control. *International Journal of Robust and Nonlinear Control*, 4(4):421–448, 1994.
- [GP16] Nicolae Gârleanu and Lasse Heje Pedersen. Dynamic portfolio choice with frictions. *Journal of Economic Theory*, 165:487–516, 2016.
- [Hau71] U. Haussmann. Optimal stationary control with state control dependent noise. *SIAM Journal on Control*, 9(2):184–198, 1971.
- [Hay14] Simon S. Haykin. *Adaptive Filter Theory*. Pearson Education, Harlow, UK, 5th edition, 2014.
- [How60] R. A. Howard. *Dynamic Programming and Markov Processes*. The MIT Press, Cambridge, MA, 1960.

- [Isi99] Alberto Isidori. *Nonlinear Control Systems II*. Springer London, 1999.
- [Iye05] Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.
- [JJ13] Z. P. Jiang and Yu Jiang. Robust adaptive dynamic programming for linear and nonlinear systems: An overview. *European Journal of Control*, 19(5):417 – 425, 2013.
- [JJ17] Yu Jiang and Z. P. Jiang. *Robust Adaptive Dynamic Programming*. Wiley-IEEE Press, Hoboken, NJ, 2017.
- [JTP94] Z. P. Jiang, Andrew R. Teel, and Laurent Praly. Small-gain theorem for ISS systems and applications. *Mathematics of Control, Signals and Systems*, 7(2):95–120, 1994.
- [KBL70] David L. Kleinman, S. Baron, and W.H. Levison. An optimal control model of human response part i: Theory and validation. *Automatica*, 6(3):357 – 369, 1970.
- [KD98] Miroslav Krstić and Hua Deng. *Stabilization of Nonlinear Uncertain Systems*. Springer-Verlag New York, 1998.
- [Kha02] H. K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2002.
- [KKK95] Miroslav Krstić, I. Kanellakopoulos, and Peter V. Kokotović. *Nonlinear and Adaptive Control Design*. John Wiley & Sons, Inc., New York, NY, 1995.
- [Kle68] David L. Kleinman. On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control*, 13(1):114–115, 1968.
- [KS72] Huibert Kwakernaak and R. Sivan. The maximally achievable accuracy of linear optimal regulators and linear optimal filters. *IEEE Transactions on Automatic Control*, 17(1):79–86, 1972.
- [Kuč73] Vladimír Kučera. A review of the matrix Riccati equation. *Kybernetika*, 9(1):42–61, 1973.
- [KY03] Harold J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer New York, 2003.
- [Lib12] D. Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, Princeton, NJ, 2012.
- [Lit15] Michael L. Littman. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521(7553):445–451, 05 2015.
- [LJH14] Tengfei Liu, Z. P. Jiang, and David J. Hill. *Nonlinear Control of Dynamic Networks*. CRC Press, New York, NY, 2014.
- [LL13] Frank L. Lewis and Derong Liu. *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. John Wiley & Sons, Inc., Piscataway, NJ, 2013.
- [LXM13] Shiao Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust Markov decision processes. In C.J.C. Burges, L. Bottou, M. Welling, Zoubin. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 701–709. Curran Associates, Inc., 2013.
- [Mer71] Robert .C Merton. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, 3(4):373 – 413, 1971.
- [NEG05] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2017/10/23 2005.
- [Pha09] Huyèn Pham. *Continuous-time Stochastic Control and Optimization with Financial Applications*. Springer-Verlag Berlin Heidelberg, 2009.
- [Pow07] Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, Inc., New York, 2007.
- [Put05] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., Hoboken, NJ, 2005.
- [PW96] Laurent Praly and Yuan Wang. Stabilization in spite of matched unmodeled dynamics and an equivalent definition of input-to-state stability. *Mathematics of Control, Signals and Systems*, 9(1):1–33, 1996.
- [RS80] M. Reed and B. Simon. *Methods of Modern Mathematical Physics: Functional analysis*. Academic Press, San Diego, 1980.
- [Sas99] Shankar Sastry. *Nonlinear Systems: Analysis, Stability, and Control*. Springer-Verlag New York, 1999.

- [SB98] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.
- [SBPW04] Jennie Si, Andrew G. Barto, Warren B. Powell, and Donald C. Wunsch, editors. *Handbook of Learning and Approximate Dynamic Programming*. Wiley-IEEE Press, Piscataway, NJ, 2004.
- [SBW92] Richard S. Sutton, Andrew G. Barto, and Ronald J. Williams. Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems*, 12(2):19–22, April 1992.
- [SH69a] A. W. Starr and Yu Chi Ho. Further properties of nonzero-sum differential games. *Journal of Optimization Theory and Applications*, 3(4):207–219, 1969.
- [SH69b] A. W. Starr and Yu Chi Ho. Nonzero-sum differential games. *Journal of Optimization Theory and Applications*, 3(3):184–206, 1969.
- [Sko56] A. V. Skorokhod. Limit theorems for stochastic processes. *Theory of Probability & Its Applications*, 1(3):261–290, 1956.
- [Son98] Eduardo D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer New York, 2nd edition, 1998.
- [Ste01] J. Michael Steele. *Stochastic Calculus and Financial Applications*. Springer New York, 2001.
- [Tao03] Gang Tao. *Adaptive Control Design and Analysis*. Wiley-IEEE Press, 2003.
- [Tod05] Emanuel Todorov. Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system. *Neural computation*, 17(5):1084–1108, 2005.
- [Tsi94] John N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202, 1994.
- [TVR97] John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.
- [TVR99] John N. Tsitsiklis and Benjamin Van Roy. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Transactions on Automatic Control*, 44(10):1840–1851, Oct 1999.
- [vdS17] Arjan van der Schaft. *L₂-Gain and Passivity Techniques in Nonlinear Control*. Springer International Publishing, 3rd edition, 2017.
- [Wil71] Jacques L. Willems. Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Transactions on Automatic Control*, 16(6):621–634, 1971.
- [Won67] W. Wonham. Optimal stationary control of a linear system with state-dependent noise. *SIAM Journal on Control*, 5(3):486–500, 1967.
- [Zam66] 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 Transactions on Automatic Control*, 11(2):228–238, Apr 1966.
- [ZL00] Xun Yu Zhou and D. Li. Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics & Optimization*, 42(1):19–33, 2000.

(T. Bian) BANK OF AMERICA MERRILL LYNCH, ONE BRYANT PARK, NEW YORK, NY 10036
E-mail address, T. Bian: tbian@nyu.edu

(Z. P. Jiang) CONTROL AND NETWORKS LAB, DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING, TANDON SCHOOL OF ENGINEERING, NEW YORK UNIVERSITY, 5 METROTECH CENTER, BROOKLYN, NY 11201
E-mail address, Z. P. Jiang: zjiang@nyu.edu