Robust Approximate Dynamic Programming and Global Stabilization with Nonlinear Dynamic Uncertainties

Yu Jiang and Zhong-Ping Jiang

***Abstract*— We propose a framework of robust approximate dynamic programming (robust-ADP), which is aimed at computing globally asymptotically stabilizing, suboptimal, control laws with robustness to dynamic uncertainties, via on-line/off-line learning. The system studied in this paper is an interconnection of a linear model with fully measurable state and unknown dynamics, and a nonlinear system with unmeasured state and unknown system order and dynamics. Differently from other ADP schemes in the past literature, the robust-ADP framework allows for learning from an unknown environment in the presence of dynamic uncertainties. The main contribution of the paper is to show that robust optimal control problems can be solved by integration of ADP and small-gain techniques.**

I. Introduction

Approximate/adaptive dynamic programming (ADP) is a methodology inspired by the learning behavior from biological systems, and it has become an effective approach for solving optimal control problems in recent years. In an ADP structure, an agent computes the optimal control policy by gradually adapting to the uncertain environment over time, i.e., it optimizes a predefined cost function by utilizing the model state and limited output information. The concept of ADP was originally developed by Werbos in [21], [22], [23], and [24]. Based on the specific ADP schemes: heuristic dynamic programming (HDP) [24] and action-dependent heuristic dynamic programming (ADHDP) [22] (or Q-learning [20]), various ADP algorithms emerged, and they have been studied both in theory and applications (see, for example, [2], [1], [18], [14], [26], [5], [6]).

A common assumption in previous ADP-based control methods is that the plant to be controlled is of a known system order, and its state is either directly measurable, or reconstructible from output measurements [14], [5]. However, the system order may be unknown because of the dynamic uncertainty. This problem, often formulated as robust control, cannot be viewed as a special case of output feedback control. Therefore, as pointed out by Werbos in [25], an important question in ADP is how to conduct learning and assure convergence using only limited information and partial-state measurements of a given system, the order of which is completely unknown. In this case, the ADP schemes developed in the past literature may fail to guarantee not only optimality, but also the stability of the closed-loop system when dynamic uncertainty occurs.

Therefore, in this paper we propose new learning strategies for ADP with robustness to dynamic uncertainties. In order to perform stability analysis for the interconnected systems, we adopt the notion of input-to-state stability (ISS) [15], [16], which has been proved to be an efficient tool for nonlinear system analysis and synthesis. Then, we develop both on-line and off-line learning strategies that compute globally asymptotically stabilizing control policies for the overall system in finite steps. We achieve the robust stability and suboptimality properties for the overall system, by means of Lyapunov and small-gain techniques [9], [7].

This paper is organized as follows. In Section 2, we formulate the control problem and introduce some tools from modern nonlinear control theory and an iterative algorithm for solving LQR problems. In Section 3, we develop both on-line and off-line robust-ADP algorithms, and prove their convergence. In Section 4, two examples are numerically simulated to illustrate the efficiency of the presented algorithms. Finally, concluding remarks are given in Section 5.

Throughout this paper, we use $\mathbb{R}_+$ and $\mathbb{Z}_+$ to denote the sets of non-negative real numbers and non-negative integers, respectively. Vertical bars $|\cdot|$ represent the Euclidean norm for vectors, or the induced matrix norm for matrices. For any piecewise continuous function u, $\|u\|$ denotes $\sup\{|u(t)|, t \geq 0\}$. We use $\otimes$ to indicate Kronecker product, and $\text{vec}(A)$ is defined to be the mn-vector formed by stacking the columns of A on top of one another, i.e., $\text{vec}(A) = [a_1^T \; a_2^T \; \cdots \; a_m^T]^T$, where $a_i \in \mathbb{R}^n$ are the columns of $A \in \mathbb{R}^{n\times m}$. I_n stands for the $n \times n$ identity matrix. A control law is also called a *policy*, and it is said to be *globally asymptotically stabilizing* if under the policy, the closed-loop system is globally asymptotically stable (GAS) at an equilibrium of interest [10].

II. Problem formulation and preliminaries

In this section, we begin with the problem formulation. Then, we recall some important tools from modern nonlinear control, and an iterative algorithm for solving linear optimal control problems. All these tools will be helpful for developing robust-ADP algorithms in the next section.

A. Problem formulation

We consider the following continuous-time system which is a linear model interconnected with nonlinear dynamic

This work has been supported in part by NSF grants DMS-0906659 and ECCS-1101401.

Y. Jiang and Z. P. Jiang are with the Control and Networks Lab, Department of Electrical and Computer Engineering, Polytechnic Institute of New York University, Brooklyn, NY 11201, USA. Z. P. Jiang is also with the College of Engineering, Beijing University, China. Email: yu.jiang@nyu.edu, zjiang@poly.edu

uncertainties, characterized by the z-system:

$$\dot{x} = Ax + B(u + \Delta(z, y)), \quad (1)$$
$$\dot{z} = g(z, y), \quad y = Cx, \quad (2)$$

where $x \in \mathbb{R}^n$ is the measured component of the state available for feedback control; $z \in \mathbb{R}^q$ is the unmeasurable part of the state with unknown order q; $u \in \mathbb{R}^m$ is the control input; $y \in \mathbb{R}^p$ is the system output; $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ are constant matrices with (A, B) controllable, (A, C) observable, and A unknown; $g : \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}^q$ and $\Delta : \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}^m$ are two unknown locally Lipschitz functions satisfying $g(0, 0) = 0$ and $\Delta(0, 0) = 0$.

Our goal is to find, if possible, a control policy that globally asymptotically stabilizes the system composed of (1) and (2), while achieving some optimality properties. Although robust adaptive and nonlinear control theory can be applied to obtain an adaptive regulator [8], the obtained adaptive controllers are often not optimal. By means of ADP, we compute the control policy through direct on-line or off-line learning in an attempt to optimize some given integral-quadratic cost function. A fundamental difference between the robust-ADP problem we address in this paper and previously introduced ADP-based feedback control problems is that we address the presence of dynamic uncertainties with unknown system order. This problem of robust-ADP-based feedback control design may seem to be a special case of output feedback ADP control. However, here we do not seek to build up a nonlinear observer to reconstruct the unmeasured state. Nonetheless, it should be noted that nonlinear observer design itself is a daunting task within the control systems community.

B. The ISS property

Consider the following control system having $x \in \mathbb{R}^n$ as the state, $u \in \mathbb{R}^m$ as the input, and $y \in \mathbb{R}^p$ as the output:

$$\dot{x} = f(x, u), \quad (3)$$
$$y = h(x, u), \quad (4)$$

where f is a locally Lipschitz function and h is a continuous function.

Definition 2.1 ([15]): The system comprised of (3)-(4) is said to be *input-to-state stable* (ISS) with gain γ if, for any measurable essentially bounded input u and any initial condition $x(0)$, the solution $x(t)$ exists for every $t \geq 0$ and satisfies

$$|x(t)| \leq \beta(|x(0)|, t) + \gamma(\|u\|), \quad (5)$$

where β, γ are of class $\mathcal{KL}$ and of class $\mathcal{K}$, respectively [10].

The following theorem gives necessary and sufficient conditions for the ISS property.

Theorem 2.1 ([16]): System (3) is ISS if and only if there exists a continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}_+$, such that the following hold for all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$:

$$\alpha_1(|x|) \leq V(x) \leq \alpha_2(|x|), \quad (6)$$
$$\tfrac{\partial V}{\partial x} f(x, u) \leq -\alpha_3(|x|) + \alpha_4(|u|), \quad (7)$$

where α_1, α_2, α_3 are of class $\mathcal{K}_\infty$, and α_4 is of class $\mathcal{K}$.

C. Linear quadratic regulator (LQR) theory

Consider the linear system

$$\dot{x} = Ax + Bu, \quad (8)$$

where A, B are the same as in (1). The objective of the problem is to find an optimal linear feedback gain $K^* \in \mathbb{R}^{m\times n}$, such that under control policy $u = -K^*x$, the following cost is minimized

$$J = \int_0^\infty \left(x^T Q x + u^T R u\right) dt, \quad (9)$$

where $Q \geq 0$, $R > 0$ are symmetric matrices, with $(A, Q^{1/2})$ observable.

According to linear optimal control theory [13], the optimal feedback gain is determined as

$$K^* = R^{-1} B^T P^*, \quad (10)$$

where the symmetric matrix $P^* > 0$ is the unique solution of the well-known algebraic Riccati equation (ARE) [13]:

$$P^* A + A^T P^* + Q - P^* B R^{-1} B^T P^* = 0. \quad (11)$$

The following Kleinman algorithm [11] gives an efficient way to numerically solve the ARE (11):

1) Choose K_0 such that $A - BK_0$ is Hurwitz and set $k = 0$.
2) Solve P_k from the following Lyapunov equation:
$$P_k A_k + A_k^T P_k + Q + K_k^T R K_k = 0 \quad (12)$$
where $A_k = A - BK_k$.
3) Update the feedback gain matrix using
$$K_{k+1} = R^{-1} B^T P_k. \quad (13)$$
4) Go to 2), and repeat until convergence is attained.

Under this algorithm, it has been shown in [11] that

1) A_k is Hurwitz.
2) $P^* \leq P_{k+1} \leq P_k$.
3) $\lim\limits_{k\to\infty} K_k = K^*$, $\lim\limits_{k\to\infty} P_k = P^*$.

III. ROBUST-ADP DESIGNS

In this section, we first investigate how optimality and stability are affected in the presence of dynamic uncertainties. Then, we develop on-line and off-line robust-ADP learning strategies to obtain globally and robustly asymptotically stabilizing control policies.

A. Optimality and asymptotic stability

To begin with, let us make a few assumptions, which are often required in the literature of nonlinear control design [12], [4], [8].

Assumption 3.1: The z-subsystem (2) is ISS with respect to y as the input.

Assumption 3.2: There exist a continuously differentiable, positive definite and radially unbounded function $W : \mathbb{R}^q \to \mathbb{R}$, and two constants $c_1 > 0$, $c_2 \geq 0$, such that

$$\frac{\partial W}{\partial z} g(z, y) \leq -c_1 |\Delta(z, y)|^2 + c_2 |y|^2, \quad (14)$$

for all $z \in \mathbb{R}^q$ and $y \in \mathbb{R}^p$.

The following lemma connects the GAS property with the selection of the weighting matrices Q and R.

Lemma 3.1: Let $u = -K^* x$ be the optimal control policy of system (8) and assume the weighting matrices in (9) satisfying $Q > \gamma C^T C$ and $R = I_m$. Then, the control policy $u = -K^* x$ globally asymptotically stabilizes (1) and (2), if $\gamma > \frac{c_2}{c_1}$.

Proof: Define $V = x^T P^* x$. Then, along the solutions of (1), we have

$$\begin{aligned} \dot{V} &= x^T \left[(A - BK^*)^T P^* + P^*(A - BK^*)\right] x \\ &\quad + x^T P^* B\Delta + \Delta^T B^T P^* x \\ &\leq -x^T Q x - |\Delta - B^T P^* x|^2 + |\Delta|^2 \\ &\leq -\gamma |y|^2 + |\Delta|^2. \end{aligned} \tag{15}$$

Setting $V_1 = V(x) + \frac{1}{c_1} W(z)$, and using Assumption 3.2, along the trajectories of (1) and (2) we have

$$\dot{V}_1 = \dot{V} + \frac{1}{c_1}\dot{W} \leq -(\gamma - \frac{c_2}{c_1})|y|^2.$$

By Assumption 3.1, all solutions of the closed-loop system are globally bounded. Moreover, using the observability of (A, C) together with Assumption 3.1, a direct application of LaSalle's Invariance Principle [10] yields the GAS property of the trivial solution of the closed-loop system. ■

The next theorem shows that the GAS property can be attained after finite steps of iterations using Kleinman's algorithm introduced in the previous section.

Remark 3.1: Note that only the ratio c_2/c_1 is necessary for the selection of an appropriate matrix Q. The condition on γ and c_2/c_1 can also be interpreted using the small-gain theorem [9] under the Lyapunov formulation [7].

Lemma 3.2: Under the conditions of Lemma 3.1, there exists a sufficiently small constant $\epsilon > 0$, such that for all symmetric matrix $P > 0$ satisfying $|P - P^*| < \epsilon$, the overall system composed of (1) and (2) is GAS under $u = -B^T P x$.

Proof: For any symmetric matrix $P > 0$, we have

$$A^T P + PA + \hat{Q} - PBB^T P = 0, \tag{16}$$

where

$$\begin{aligned} \hat{Q} &= Q + (P^* - P)A + A^T(P^* - P) \\ &\quad + PBB^T P - P^* BB^T P^*. \end{aligned}$$

According to the conditions of Lemma 3.1, there exists a constant $\alpha > 0$, such that $Q - \gamma C^T C > \alpha I_n$. Then, by continuity, there exists $\epsilon > 0$, such that for any symmetric matrix $P > 0$ satisfying $|P - P^*| < \epsilon$, we have $\hat{Q} > Q - \alpha I_n$, which implies $\hat{Q} > \gamma C^T C$. Therefore, by Lemma 3.1, the control policy $u = -B^T P x$ globally asymptotically stabilizes (1) and (2). ■

For simplicity, in the remainder of this paper we assume $R = I_m$, leaving the matrix Q to be designed.

B. On-line learning strategy

Now, we investigate how to compute a globally asymptotically stabilizing control policy on-line. We first consider solving the equation (12) using the measurable model state and the output information about the dynamic uncertainties.

For this purpose, let us make the following assumption.

Assumption 3.3: The output, $\Delta(z, y)$, of the dynamic uncertainty, is assumed to be available on some disjoint intervals

$$\bigcup_{j=0}^{\infty} [t_j, t_j + \delta t] \bigcap [0,\ t], \tag{17}$$

where $t \geq 0$ denotes the current time, $\delta t > 0$ is a positive constant, $\{t_j\}$ is an increasing sequence satisfying $0 \leq t_j < t_j + \delta t < t_{j+1}$ for all $j \in \mathbb{Z}_+$.

Next, we show that given K_k such that $A - BK_k$ is Hurwitz, it is possible to solve P_k from (12) using the available information of x and Δ, instead of the matrix A.

Let $v = -K_k x$ and $u = v + e$, where e is the *exploration* noise to be determined later. Then, for any $t \geq 0$, along the trajectories of (1), we have

$$\begin{aligned} & x^T(t + \delta t) P_k x(t + \delta t) - x^T(t) P_k x(t) \\ &= \int_t^{t+\delta t} \left[x^T(A_k^T P_k + P_k A_k)x + 2x^T P_k B \hat{\Delta}\right] d\tau \\ &= \int_t^{t+\delta t} \left(-x^T Q x - |v|^2 + 2x^T P_k B \hat{\Delta}\right) d\tau, \end{aligned} \tag{18}$$

where $A_k = A - BK_k$, $\hat{\Delta} = \Delta + e$.

Applying Kronecker product representation [3] gives

$$x^T P_k x = (x^T \otimes x^T)\mathrm{vec}(P_k), \tag{19}$$

$$x^T P_k B \hat{\Delta} = (\hat{\Delta}^T \otimes x^T)(B^T \otimes I_n)\mathrm{vec}(P_k). \tag{20}$$

Therefore, (18) is equivalent to

$$\begin{aligned} & [x^T(t) \otimes x^T(t) - x^T(t + \delta t) \otimes x^T(t + \delta t) \\ & +2 \int_t^{t+\delta t} (\hat{\Delta}^T \otimes x^T) d\tau \ (B^T \otimes I_n)]\mathrm{vec}(P) \\ = & \int_t^{t+\delta t} \left(x^T Q x + |v|^2\right) d\tau. \end{aligned}$$

Furthermore, define

$$\Phi_k = \begin{bmatrix} \Delta x(t_0^{(k)}) + 2I_{\Delta x}(t_0^{(k)})(B^T \otimes I_n) \\ \Delta x(t_1^{(k)}) + 2I_{\Delta x}(t_1^{(k)})(B^T \otimes I_n) \\ \vdots \\ \Delta x(t_{l-1}^{(k)}) + 2I_{\Delta x}(t_{l-1}^{(k)})(B^T \otimes I_n) \end{bmatrix}, \quad \Psi_k = \begin{bmatrix} c(t_0^{(k)}) \\ c(t_1^{(k)}) \\ \vdots \\ c(t_{l-1}^{(k)}) \end{bmatrix},$$

$$\begin{aligned} \Delta x(t) &= x^T(t) \otimes x^T(t) - x^T(t + \delta t) \otimes x^T(t + \delta t), \\ I_{\Delta x}(t) &= \int_t^{t+\delta t} \left(\hat{\Delta}^T \otimes x^T\right) d\tau, \\ c(t) &= \int_t^{t+\delta t} \left(x^T Q x + |v|^2\right) d\tau, \end{aligned}$$

where l can be any integer satisfying $l \geq n^2$, $\{t_j^{(k)}\}_{j=0}^{l-1}$ is a subset of $\{t_j\}$ defined in Assumption 3.3, the superscript (k)

denotes that on intervals $[t_j^{(k)}, t_j^{(k)} + \delta t]$ for $j = 0, 1, \cdots, l-1$, the control policy applied to the system is $u = -K_k x + e$.

Consequently, (18) implies the following algebraic linear equations:

$$\Phi_k \text{vec}(P_k) = \Psi_k \tag{21}$$

Next, to guarantee the uniqueness of P_k in (21), we make the following assumption.

Assumption 3.4: Φ_k has full column rank for all $k \in \mathbb{Z}_+$, i.e., $\text{rank}(\Phi_k) = n^2$.

Remark 3.2: A practical way to satisfy Assumption 3.4 is by choosing the appropriate exploration noise $e(t)$. The exploration noise can be random noise as used in [2], [1]. In this paper, we construct $e(t)$ by adding sinusoidal functions with different frequencies as in [5].

Under Assumption 3.4, $\text{vec}(P_k)$ can be solved from

$$\text{vec}(P_k) = (\Phi_k^T \Phi_k)^{-1} \Phi_k^T \Psi_k. \tag{22}$$

So far, we have developed a way that solves the Lyapunov equation (12) using the information of x and Δ obtained online, instead of the knowledge of A. The equivalence between (12) and (21) is shown in the following lemma.

Lemma 3.3: Under Assumption 3.4, the solutions P_k of (12) and (21) are the same.

Proof: From the above derivations, we see that for any given K_k such that $A - BK_k$ is Hurwitz, the solution P_k of (12) satisfies (21). Under Assumption 3.4, the solution of (21) is unique. On the other hand, any solution of (21) also satisfies (12), and we know the solution of (12) is unique due to the fact that $A - BK_k$ is Hurwitz. Therefore, the solutions of (12) and (21) are thus the same. ■

Now, we give the on-line robust-ADP algorithm:

Algorithm 3.1: (The on-line robust-ADP Algorithm)

1) Choose K_0 such that $A - BK_0$ is Hurwitz. Set $\gamma > \frac{c_2}{c_1}$, a sufficiently small constant $\epsilon > 0$, and let $k = 0$.
2) Apply the control policy $u = -K_k x + e$ to the system composed of (1) and (2). Solve P_k from (22).
3) Update the control policy by obtaining a new feedback gain matrix K_{k+1} using (13).
4) Go to 5) if $|P_{k+1} - P_k| \leq \epsilon$. Otherwise, set $k = k + 1$ and go to 2).
5) Apply the control policy $u = -K_k x$ to (1) and (2).

Remark 3.3: By Lemma 3.2 and Lemma 3.3, there exists a constant $\epsilon' > 0$, such that $|P_k - P^*| < \epsilon'$ implies $u_k = -B^T P_k x$ is a globally and robustly asymptotically stabilizing control policy. By [11], we can always find a sufficiently small constant $\epsilon > 0$ such that the following implication holds

$$|P_k - P_{k+1}| \leq \epsilon \;\Rightarrow\; |P_k - P^*| \leq \epsilon'. \tag{23}$$

Hence, under the control policy $u = -B^T P_k x$ applied in Step 5), the system composed of (1) and (2) is GAS.
The following theorem summarizes the convergence property of the on-line robust-ADP algorithm leading to a GAS-stabilizing controller with suboptimality properties.

Theorem 3.1: Under Assumptions 3.1, 3.2, 3.3 and 3.4, Algorithm 3.1 gives a control policy that globally and robustly asymptotically stabilizes (1)-(2) and achieves some suboptimality property.

C. *Off-line learning strategy*

In a more general setting, the output $w(z)$ of dynamic uncertain z-system is not available at all, hence Assumption 3.3 may not hold. To avoid this obstacle, we develop an off-line robust-ADP strategy, which computes a globally asymptotically stabilizing control policy only using the input and state information of (8) on $[0, T]$ where T can be any positive number.

Consider $u = v$ to be the input of (8) on the interval $[0, T]$, and let the corresponding solution be $x_v(t)$. Our goal is to compute a globally asymptotically stabilizing control law based on $v(t)$ and $x_v(t)$. To this end, we first randomly select a sequence $\{t_j\}_{j=0}^{l}$, such that $t_0 = 0$, $t_l = T$, and $t_j < t_{j+1}$ for $j = 0, 1, \cdots, l-1$.

Next, notice that for any stabilizing control policy $u_k = -K_k x_v$, we can rewrite (8) as

$$\dot{x}_v = Ax_v + Bu_k + B(v - u_k), \tag{24}$$

where u_k is regarded as the *virtual* input to the system, and $v - u_k$ is treated as the measurable disturbance input.

Now, along the trajectories of (24), it follows that

$$\begin{aligned} & x_v(t_j)^T P_k x_v(t_j) - x_v(t_{j+1})^T P_k x_v(t_{j+1}) \\ = & \int_{t_j}^{t_{j+1}} x_v^T (Q + K_k^T K_k) x_v d\tau \\ & -2 \int_{t_j}^{t_{j+1}} x_v^T P_k B(v + K_k x_v) d\tau. \end{aligned} \tag{25}$$

Using Kronecker product representation, we have

$$\begin{aligned} & \left[x_v^T(t_j) \otimes x_v^T(t_j) - x_v^T(t_{j+1}) \otimes x_v^T(t_{j+1})\right] \text{vec}(P_k) \\ = & \int_{t_j}^{t_{j+1}} (x_v^T \otimes x_v^T) \text{vec}(Q + K_k^T K_k) d\tau \\ & -2 \int_{t_j}^{t_{j+1}} (v^T \otimes x_v^T)(B^T \otimes I_n) \text{vec}(P_k) d\tau \\ & -2 \int_{t_j}^{t_{j+1}} (x_v^T \otimes x_v^T)(K_k^T B^T \otimes I_n) \text{vec}(P_k) d\tau. \end{aligned} \tag{26}$$

Further, define

$$\Theta_k = \begin{bmatrix} \delta x^1 + 2I_{vx}^1 (B^T \otimes I_n) + 2I_{xx}^1 (K_k^T B^T \otimes I_n) \\ \delta x^2 + 2I_{vx}^2 (B^T \otimes I_n) + 2I_{xx}^2 (K_k^T B^T \otimes I_n) \\ \vdots \\ \delta x^l + 2I_{vx}^l (B^T \otimes I_n) + 2I_{xx}^l (K_k^T B^T \otimes I_n) \end{bmatrix}_{l \times n^2},$$

$$\Xi_k = \begin{bmatrix} I_{xx}^1 \\ I_{xx}^2 \\ \vdots \\ I_{xx}^l \end{bmatrix}_{l \times n^2} \text{vec}(Q + K_k^T K_k),$$

where, for all $j = 1, 2, \cdots, l$,

$$\begin{aligned}\delta x^j &= x_v^T(t_{j-1}) \otimes x_v^T(t_{j-1}) - x_v^T(t_j) \otimes x_v^T(t_j),\\ I_{vx}^j &= \int_{t_{j-1}}^{t_j} (v^T \otimes x_v^T) d\tau, \quad I_{xx}^j = \int_{t_{j-1}}^{t_j} (x_v^T \otimes x_v^T) d\tau.\end{aligned}$$

Then, (26) implies the following matrix form of linear equations:

$$\Theta_k \text{vec}(P_k) = \Xi_k. \tag{27}$$

Assumption 3.5: Θ_k has full column rank, for all $k \in \mathbb{Z}_+$.

Remark 3.4: Similar to the on-line learning case, to assure Assumption 3.5 holds, we require that $v(t)$ on the interval $[0, T]$ contain enough different frequencies.

Algorithm 3.2: (The off-line robust-ADP Algorithm)

1) Choose K_0 such that $A - BK_0$ is Hurwitz. Set $\gamma > \frac{c_2}{c_1}$, sufficiently small $\epsilon > 0$, and let $k = 0$.
2) Solve P_k from (27)
3) Update the virtual control policy using (13).
4) Stop, if $|P_{k+1} - P_k| \leq \epsilon$. Otherwise, set $k = k + 1$ and go to 2).

The next theorem summarizes the convergence property of the off-line Algorithm 3.2 that results in a GAS-stabilizing controller with suboptimality properties.

Theorem 3.2: Under Assumption 3.5, the off-line robust-ADP algorithm computes a control policy that globally and robustly asymptotically stabilizes (1) and (2).

Remark 3.5: Compared with the continuous-time ADP method in [18], the off-line robust-ADP algorithm has two distinctive features. First, only the information of the input and partial state over a finite interval is necessary for learning. Second, there is no need to alter the control input on-line to get new input and state information. Instead, we decompose the actual input into two parts: the virtual control and the disturbance. Then, we use both of them to compute the corresponding cost matrix P_k, based on which P_k we update the virtual control. In this way, there is no need for us to change the actual input to the system, and we are able to use repeatedly the same input and partial state information for each iteration until convergence is attained.

IV. Applications

In this section, we apply the robust-ADP algorithms to solve two examples. The first one is a second-order linear system interconnected with a scalar nonlinear system. The second example comes from the load-frequency controller design for power systems. The on-line and off-line algorithms will be applied to the examples, respectively.

A. Example 1

Consider the following system,

$$\dot{x} = Ax + \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} (u + \Delta(z)), \quad \dot{z} = g(z, x) \tag{28}$$

where $x = [x_1, \ x_2]^T$ is the measurable model state, z is not available for feedback control, A is an unknown matrix but its eigenvalues have negative real parts. The z-subsystem is assumed to satisfy both Assumptions 3.1 and 3.2 with $\frac{c_2}{c_1} < 2$. For all $t \geq 0$, Δ is only available on $\bigcup_{j=0}^{\infty} [j, j + 0.2] \bigcap [0, \ t]$.

In order to apply the on-line robust-ADP algorithm, we start from $K_0 = [0, \ 0]$, and update the control policy every $10s$. The weighting matrix is set to be $Q = 2I_2$. The iteration stops whenever $|P_{k+1} - P_k| < 10^{-4}$. To assure Assumption 3.4 holds, we set the exploration noise to be $e(t) = \frac{1}{100} \sum_{i=1}^{4} [\sin(\frac{2i-1}{100} t)]$.

For the purpose of simulation, we set $g(z, x) = -z^3 + |x_1|^{3/2}$, $\Delta = z^2$, $A = \begin{bmatrix} -0.10 & -0.02 \\ 0.01 & -0.02 \end{bmatrix}$. The initial conditions are $z(0) = 10$, $x_1(0) = 6$, $x_2(0) = -7$.

Define $W = \frac{1}{2}z^2$. Then, it can be easily checked that $\frac{c_2}{c_1} = 1 < 2$ using Young's inequality.

After applying the on-line robust-ADP algorithm, the closed-loop system trajectories are shown in Figure 1.

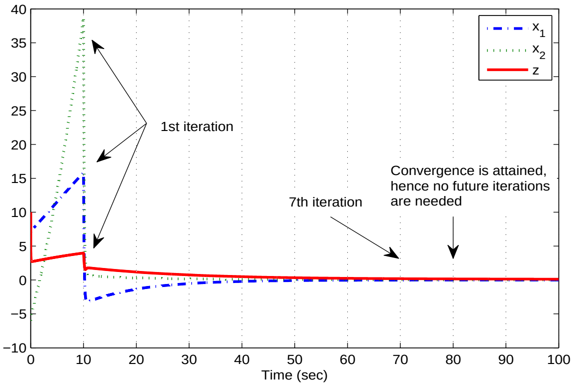

Fig. 1. Profile of the system trajectories under robust-ADP.

Initially, the overall system is not stable, but it becomes GAS after seven iterations. The final solution P_7 we have obtained via on-line learning algorithm and its difference with the optimal value is shown as follows:

$$P_7 - P^* = \begin{bmatrix} -0.1247 & -0.0042 \\ -0.0042 & -0.0237 \end{bmatrix} \times 10^{-5}.$$

B. Example 2

Consider the problem of power system load-frequency control [19], [18]:

$$\begin{aligned}\Delta \dot{P}_g &= -\frac{1}{T_T} \Delta P_g + \frac{1}{T_T} \Delta X_g, & (29)\\ \Delta \dot{X}_g &= -\frac{1}{T_G} \Delta X_g + \frac{1}{T_G} u - \frac{1}{R_0 T_G} \Delta f, & (30)\\ \Delta \dot{f} &= -\frac{1}{T_p} \Delta f + \frac{K_P}{T_P} P_g, & (31)\end{aligned}$$

where Δf is the incremental frequency deviation, ΔP_g is the incremental change in generator output, and ΔX_g is the incremental change in governor valve position. Parameters

T_g, T_t, T_p, K_p and R_0 denote the governor time constant, the turbine time constant, the plant model time constant, the plant gain and the speed regulation due to governor action, respectively. For simplicity, we do not consider the integral control of $\Delta f(t)$.

In [18], an on-line ADP algorithm is applied to formulate a state-feedback optimal control policy of this system. Here we assume that $\Delta f(t)$ is not available for feedback design. Hence, the problem cannot be solved using previous ADP methods developed in the past literatures. Now we use the proposed off-line robust-ADP to solve this problem, in the sense that we compute a control policy using ΔP_g and ΔX_g only, to achieve GAS of the closed-loop system. Note that system (29)-(31) is already in the form of (1) and (2), with $x = [\Delta P_g \ \ \Delta X_g]^T$, $z = \Delta f$, $\Delta = -\frac{1}{R_0}z$. Assumptions 3.1 and 3.2 of the Δf-subsystem can be easily checked.

After the parameters are all set, we input the following signal to the system

$$v(t) = \sin(t) + \sin(2t) + \sin(7t) + \sin(11t), \quad t \in [0, 0.1].$$

and we record the state trajectory $x(t)$ on $0 \leq t \leq 0.1$.

Finally, we choose $t_j = \frac{j}{100}$, for $j = 0, 1, \cdots, 100$, $Q = 10^3 I_2$ and run Algorithm 3.2. Similar as in [18], we set the initial control policy to be $K_0 = [0, \ 0]$.

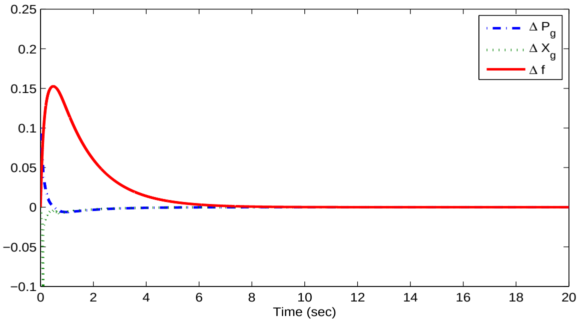

Fig. 2. Profile of the system trajectories under robust-ADP.

For simulation purpose, we set $T_T = 0.21$, $T_p = 15.0376$, $K_p = 180.4511$, $T_g = 0.0728$, and $R_0 = 0.4579$. The initial conditions $\Delta f(0) = 0$, $\Delta P_g(0) = 0.1$, and $\Delta X_g(0) = 0$ are taken from [18]. The iteration stops when the criterion $|P_{k+1} - P_k| \leq 10^{-2}$ is satisfied, after updating the policy for 10 times. The final solution and its optimal value are shown as follows:

$$P_8 = \begin{bmatrix} 87.3849 & 0.9429 \\ 0.9429 & 2.2409 \end{bmatrix}, \quad P^* = \begin{bmatrix} 87.3831 & 0.9429 \\ 0.9429 & 2.2408 \end{bmatrix}.$$

Under the control policy formulated by the proposed off-line robust-ADP algorithm, the closed-loop system is GAS as illustrated in Figure 2.

V. Conclusions and Future Work

In this paper, a new framework of robust-ADP has been proposed for nonlinear control systems with dynamic uncertainties. This novel methodology is developed by integration of ADP [21], and the tools related to ISS and small-gain theories from nonlinear control theory [9], [7], [15], [16]. In this paper, a first step has been made to show that ADP can be used to handle robust optimal control problems. Because of the fact that robust-ADP allows significant uncertainties in systems, it potentially has numerous real-world applications.

References

[1] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, pp. 473-481, 2007.

[2] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," *Proceedings of American Control Conference*, vol. 3, pp. 3475-3479, 1994.

[3] R. A. Horn and C. R. Johnson, *Matrix Analysis*, NY: Cambridge University Press, 1985.

[4] A. Isidori, *Nonlinear Control Systems. vol. II*, Springer-Verlag, 1999.

[5] Y. Jiang and Z. P. Jiang, "Approximate dynamic programming for output feedback control," *Proceedings of the 29th Chinese Control Conference*, pp. 5815-5820, 2010.

[6] Y. Jiang and Z. P. Jiang, "Approximate dynamic programming for optimal stationary control with control-dependent noise," *IEEE Transactions on Neural Networks*, in press.

[7] Z. P. Jiang, I. Mareels and Y. Wang, "A Lyapunov formulation of the nonlinear small gain theorem for interconnected ISS systems," *Automatica*, vol. 32, no. 8, pp. 1211-1215, 1996.

[8] Z. P. Jiang and L. Praly, "Design of robust adaptive controllers for nonlinear systems with dynamic uncertainties," *Automatica*, vol. 34, no. 7, pp. 825-840, 1998.

[9] Z. P. Jiang, A. R. Teel, and L. Praly, "Small-gain theorem for ISS systems and applications," *Mathematics of Control, Signals, and Systems*, vol. 7, no. 2, pp. 95-120, 1994.

[10] H. K. Khalil, *Nonlinear Systems* (3rd edition), Prentice Hall, 2002.

[11] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114-115, 1969.

[12] M. Krstic, I. Kanellakopoulos and P. V. Kokotovic, *Nonlinear and Adaptive Control Design*, John Wiley, 1995.

[13] F. L. Lewis and V. L. Syrmos, *Optimal Control*, Wiley, 1995.

[14] F. L. Lewis, K. G. Vamvoudakis, "Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data," *IEEE Transactions Systems, Man, and Cybernetics, Part B*, vol. 41, no. 1, pp. 14-23, 2011.

[15] E. D. Sontag, "Further facts about input to state stabilization," *IEEE Transactions on Automatic Control*, vol. 35, no. 4, pp. 473-476, 1990.

[16] E. D. Sontag and Y. Wang, "On characterizations of the input-to-state stability property," *Systems & Control Letters*, vol. 24, pp. 351-359, 1995.

[17] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[18] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477-484, 2009.

[19] Y. Wang, R. Zhou, and C. Wen, "Robust load-frequency controller design for power systems," *IEE Proceedings-Generation, Transmission and Distribution*, vol. 104, no. 1, pp. 11-16, 1993.

[20] C. Watkins, *Learning from delayed rewards*, PhD thesis, King's College of Cambridge, UK, 1989.

[21] P. J. Werbos, *Beyond regression: New tools for prediction and analysis in the behavioural sciences*, Ph.D. Thesis, Harvard University, 1972.

[22] P. J. Werbos, "Neural networks for control and system identification," *Proceedings of IEEE Conference on Decision and Control*, pp. 260-265, 1989.

[23] P. J. Werbos, "A menu of designs for reinforcement learning over time," *Neural Networks for Control*, pp. 67-95, ed. W. T. Miller, R. S. Sutton, P. J. Werbos, Cambridge: MIT Press, 1991.

[24] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 493-525, 1992.

[25] P. J. Werbos, "Intelligence in the brain: A theory of how it works and how to build it," *Neural Networks*, vol. 22, no. 3, pp. 200-212, 2009.

[26] H. Zhang, Q. Wei, and D. Liu, "An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games," *Automatica*, vol. 47, no. 1, pp. 207-214, 2011.