Robust adaptive dynamic programming for linear and nonlinear systems: An overview

Zhong-Ping Jiang*, Yu Jiang

Department of Electrical and Computer Engineering, Polytechnic Institute of New York University, Brooklyn, NY 11201, USA

1. Introduction

1.1. Background

Approximate/adaptive dynamic programming (for short, ADP) is a biologically-inspired, non-model-based, computational method that has been used to compute optimal control laws; see, e.g., [43,49,62,64,66] and numerous references therein. It is well-known that conventional dynamic programming [3] requires the perfect knowledge of system dynamics and suffers from the curse of dimensionality. To avoid these difficulties, Werbos first pointed out in [61] that adaptive approximation to the Hamilton–Jacobi–Bellman (HJB) equation [37] can be achieved by designing appropriate reinforcement learning systems (see, [53] for an excellent introduction to the theory of reinforcement learning). In his seminal work [63,65,66], Werbos further proposed two basic approaches for implementing ADP: heuristic dynamic programming (HDP) and dual dynamic programming. They can be used to approximate the optimal cost function or its gradient, and their generalized versions can be found in [66] in which the approximation of the optimal control policy is considered. Similar problems were also studied by Bertsekas and Tsitsiklis [5] under the name of neuro-dynamic programming and were restricted exclusively to discrete-time systems. A rigorous development of the mathematical principles behind neuro-dynamic programming is provided, along with numerous methods and applications.

The development of ADP theory consists of three phases. In the first phase, ADP was extensively investigated within the communities of computer science and operations research. Two basic algorithms, policy iteration [17] and value iteration [3], are usually employed. In [52], Sutton introduced the temporal difference method. In 1989, Watkins proposed the well-known Q-learning method in his PhD thesis [60]. Q-learning shares similar features with the action-dependent HDP scheme proposed by Werbos in [64]. Other related research work under a discrete time and discrete state-space Markov decision process framework can be found in [4,5,7,9,43,44,45,53,54] and reference therein. In the second phase, stability is brought into the context of ADP theory while real-time control problems are studied for dynamic systems. To the best of the authors’ knowledge, Lewis is the first who has contributed to the integration of stability theory and ADP theory [38]. An essential advantage of ADP theory is that an optimal control policy can be obtained via a recursive numerical algorithm using online information without solving the HJB equation (for nonlinear systems) and the algebraic Riccati equation (ARE) (for linear systems), even when the system dynamics are not precisely known. Optimal feedback control designs for linear and nonlinear dynamic systems have been proposed by several researchers over the past few years; see, e.g., [6,8,11,16,42,56,58,59,68,69]. While most of the previous work on ADP theory was devoted to discrete-time systems (see [36] and references therein), there has been relatively less research for the continuous-time counterpart. This is mainly because ADP is considerably more difficult for continuous-time systems than for discrete-time systems. Indeed, many results

*Corresponding author
E-mail addresses: zjiang@poly.edu, zpj201@nyu.edu (Z.-P. Jiang), yu.jiang@nyu.edu (Y. Jiang).
developed for discrete-time systems [39] cannot be extended straightforwardly to continuous-time systems. Nonetheless, early attempts were made to apply Q-learning for continuous-time systems via discretization technique [2,12]. However, the convergence and stability analysis of these schemes are challenging. In [42], Murray et al. proposed an implementation method which requires the measurements of the derivatives of the state variables. As said previously, Lewis and his co-worker proposed the first solution to stability analysis and convergence proofs for ADP-based control systems by means of LQR theory [58].

A synchronous policy iteration scheme was also presented in [55]. For continuous-time linear systems, the partial knowledge of the system dynamics (i.e., the input matrix) must be precisely known. This restriction has been completely removed in [21]. A nonlinear variant of this method can be found in [27].

The third phase in the development of ADP theory is related to extensions of previous ADP results to nonlinear uncertain systems. Neural networks and game theory are utilized to address the presence of uncertainty and nonlinearity in control systems. See, e.g., [56,57,69,36]. An implicit assumption in these papers is that the system order is known and that the uncertainty is static, not dynamic. The presence of dynamic uncertainty has not been systematically addressed in the literature of ADP. By dynamic uncertainty, we refer to the mismatch between the nominal model and the real plant when the order of the nominal model is lower than the order of the real system. A closely related topic of research is how to account for the effect of unseen variables [67]. It is quite common that the full-state information is often missing in many engineering and biological applications and only the output measurement or partial-state measurements are available. Adaptation of the existing ADP theory to this practical scenario is important yet non-trivial. Neural networks are sought for addressing the state estimation problem [13,32]. However, the stability analysis of the estimator/controller augmented system is by no means easy, because the total system is highly interconnected. The configuration of a standard ADP-based control system is shown in Fig. 1.

Our recent work [20,22–25] on the development of robust variants of ADP theory is exactly targeted at addressing these challenges.

### 2. What is RADP?

RADP is developed to address the presence of dynamic uncertainty in linear and nonlinear dynamical systems. See Fig. 2 for an illustration. There are several reasons for which we pursue a new framework for RADP. First and foremost, it is well-known that building an exact mathematical model for physical systems often is a hard task. Also, even if the exact mathematical model can be obtained for some particular engineering and biological applications, simplified models are often more preferable for system analysis and control synthesis than the original complex system model. While we refer the mismatch between the simplified model and the original system to as dynamic uncertainty here, the engineering literature often uses the term of unmodeled dynamics instead. Secondly, the observation errors may often be captured by dynamic uncertainty. From the literature of modern nonlinear control [34,28,29], it is known that the presence of dynamic uncertainty makes the feedback control problem extremely challenging in the context of nonlinear systems. In order to broaden the application scope of ADP theory in the presence of dynamic uncertainty, our strategy is to integrate tools from nonlinear control theory, such as Lyapunov designs [34,19,31], input-to-state stability theory [51], and nonlinear small-gain techniques [30,28]. This way RADP becomes applicable to wide classes of uncertain dynamic systems with incomplete state information and unknown system order/dynamics.

Additionally, RADP can be applied to large-scale dynamic systems as shown in our recent paper [23]. By integrating a simple version of the cyclic-small-gain theorem [40], asymptotic stability can be achieved by assigning appropriate weighting matrices for each subsystem. Further, certain suboptimality property can be obtained. Because of several emerging applications of practical importance such as smart electric grid, intelligent transportation systems and groups of mobile autonomous agents, this topic deserves further investigations from a RADP point of view. The existence of unknown parameters and/or dynamic uncertainties, and the limited information of state variables, give rise to challenges for the decentralized or distributed controller design of large-scale systems.

### 2. ADP for completely unknown, continuous-time, linear systems

We begin with a linear time-invariant (LTI), continuous-time system

\[
\dot{x} = Ax + Bu
\]

where \(x \in \mathbb{R}^n\) is the state vector, \(u \in \mathbb{R}^m\) is the control input, \(A \in \mathbb{R}^{n \times n}\) and \(B \in \mathbb{R}^{n \times m}\) are unknown constant matrices, that are stabilizable. It is assumed that there is a known constant matrix \(K_0\) such that \(A-BK_0\) is a Hurwitz matrix.

The control objective is to find, if possible, an online, optimal control policy \(u^* = -Kx\) such that the following integral-quadratic
cost is minimized:

\[ J = \int_0^\infty (x^T Q x + u^T R u) \, dt \]  

(2)

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

A prior solution to this problem was presented in [58] under the assumption that \( B \) is a known matrix although \( A \) is unknown. We show that this assumption can be removed.

To begin with, we select a stabilizing initial gain matrix \( K_0 \) such that \( A-BK_0 \) is Hurwitz. Next, we apply \( u_0 = -K_0 x + e \) as the control input with \( e \) an exploration noise, and record the state and input information on \([t_i, t_{i+1}]\) with \( i = 0, 1, \ldots, \ell-1 \) where \( \ell > 0 \) is a sufficiently large integer. These intervals \([t_i, t_{i+1}]\) are tied to learning and information collection. Then, letting \( K_{k+1} = K^R B^TP_k \), by means of standard LQR theory and the iterative algorithm in [33], it holds

\[ \frac{d}{dt}(x^T P_k x) = x^T (P_k (A-BK_k) + (A-BK_k)^T P_k)x + 2x^T P_k B(u_0 + K_k x) \]

\[ = x^T (Q + K^T R K)x + 2(u_0 + K_k x)^T R K_{k+1} x \]

Consequently, for \( k = 0, 1, 2, \ldots \), the following iterative equation can be derived (see [21]):

\[ x^T P_k x_{k+1} = \int_{t_k}^{t_{k+1}} -x^T (Q + K^T R K)x + 2(u_0 + K_k x)^T R K_{k+1} x \, dt \]  

(3)

The existence and uniqueness of solution \( (P_k, K_{k+1}) \) to the above Eq. (4) relies on the fact that the exploration noise \( e \) is persistently exciting (PE), a standard condition for parameter estimation in adaptive control [15,18].

The following result shows that these sequences \( (P_k) \) and \( (K_k) \) converge to their optimal solutions \( P^* \) and \( K^* \), respectively, noting that \( P^* \) corresponds to the solution to the ARE equation (5) in LQR theory.

**Theorem 1** (Jiang and Jiang [21]). Under the PE condition on \( e(t) \), assume that \( P_k = P_k^\ell \) and \( K_{k+1} \) can be uniquely solved from (4), for all \( k = 0, 1, 2, \ldots \). Then, \( \lim_{k \to \infty} P_k = P^* \), \( \lim_{k \to \infty} K_k = K^* \), where \( K^* = K^R B^TP^* \), and \( P^* > 0 \) is the symmetric solution of the following ARE:

\[ A^T P + PA + Q - PB R^{-1} B^T P = 0. \]

A practical online learning algorithm is summarized as follows.

Throughout the paper, the single bars \(| \cdot |\) denote the Euclidean norm for vectors and its corresponding norm for matrices.

Online Policy Iteration Algorithm:

1. Let \( k = 0 \).
2. Solve \( P_k \) and \( K_{k+1} \) from (4).
3. Let \( k = k+1 \) and repeat Step 2 until \( |P_k - P_{k-1}| \leq \epsilon \) for \( k \geq 1 \), where the constant \( \epsilon > 0 \) can be any predefined small threshold.
4. Finally, use \( u = -K_k x \) as the approximated optimal control policy.

**Remark 2.** In order to satisfy the PE condition in [21], the exploration noise \( e \), which may be a random noise or a signal comprised of the sum of sinusoidal signals with different frequencies, is used. In addition, the number of time intervals for the collection of online data should be sufficiently large. See [21] for more detailed analysis on the PE condition.

3. RADP for partially linear composite systems

The purpose of this section is to show that Theorem 1 can be generalized to a class of uncertain partially linear composite systems, which was widely studied in the literature of nonlinear control [46]. Any system in this class can be regarded as an interconnection of a linear subsystem and a nonlinear subsystem (or dynamic uncertainty):

\[ \dot{w} = q(w, y), \]

\[ \dot{x} = Ax + B[u + \Delta(w, y)], \]

\[ y = Cx \]

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

To study the robust stabilization problem of (6)-(8), let us consider the following control system having \( x \in \mathbb{R}^n \) as the state, \( u \in \mathbb{R}^m \) as the input, and \( y \in \mathbb{R}^q \) as the output:

\[ \dot{x} = f(x, u), \quad y = h(x, u) \]

where \( f \) is a locally Lipschitz function and \( h \) is a continuous function. The following definitions are taken from [30]. Also see [51].

**Definition 3.** System (9) is said to be input-to-output stable (IOS) with gain \( \gamma \) if, for any measurable locally essentially bounded input \( u \) and any initial condition \( x(0) \), the solution \( x(t) \) exists for every \( t \geq 0 \) and satisfies

\[ |y(t)| \leq \gamma |x(0)| + \gamma \|u\| \]

(10)

where \( \beta, \gamma \) are of class \( K_{\ell} \) and of class \( K \), respectively.

**Definition 4.** System (9) is said to have the strong unboundedness observability (SUO) property with zero offset, if a function \( \beta_0 \) of class \( K_{\ell} \), a function \( \beta \) of class \( K \) exist such that, for each measurable control \( u(t) \) defined on \([0, T] \) with \( 0 < T \leq \infty \), the solution \( x(t) \) of (9) right maximally defined on \([0, T] \( 0 < T \leq \infty \)) satisfies

\[ |x(t)| \leq \beta_0(|x(0)| + |y(\|u\|)|), \quad \forall t \in [0, T]. \]

(11)

In order to achieve global asymptotic stability, let us make a few assumptions about (6), which are often required in the literature of nonlinear control design [19].

**Assumption 5.** The \( w \)-subsystem has SUO property with zero offset and is IOS with respect to \( y \) as the input and \( \Delta \) as the output.

**Assumption 6.** There exist a continuously differentiable, positive definite, radially unbounded function \( W: \mathbb{R}^{m+n} \to \mathbb{R}_+ \), and two constants \( c_1 > 0, c_2 \geq 0 \) such that

\[ \frac{dW(w)}{dt} = q(w, y) + c_1 |\Delta|^2 + c_2 |y|^2 \]

(12)

for all \( w \in \mathbb{R}^{m+n} \) and \( y \in \mathbb{R}^q \).

Online learning is conducted following the Algorithm presented in Section 2 with \( u_0 \) replaced by \( u_0 + \Delta \).

**Theorem 7** (Jiang and Jiang [22]). Under Assumptions 5 and 6, suppose the exploration noise \( e(t) \) satisfies the PE condition such that \( P_k = P_k^\ell \) and \( K_{k+1} \) can be uniquely solved from (4), for all \( k = 0, 1, 2, \ldots \). Also, let \( Q \geq c_2/C^T C \) and \( R = I_m \). Then, we have \( \lim_{k \to \infty} P_k = P^* \), \( \lim_{k \to \infty} K_k = K^* \), and \( u = -K^* x \) robustly and globally...
4. RADP for higher-dimensional systems with unmatched dynamic uncertainties

The disturbance $\Delta$ in (7) satisfies the so-called strict matching condition, because it is in the span of the input space. Now, we are ready to relax this assumption into the unmatched case. To this end, consider the following interconnected system

$$w = q(w, y),$$

$$\dot{x} = Ax + B_2 z + \Delta_1 (w, y),$$

$$\dot{z} = Ex + F_2 z + G_2 u + \Delta_2 (w, y),$$

$$y = C x$$

where $[x^T, z^T]^T \in \mathbb{R}^{n+m}$ is the vector of system states; $A \in \mathbb{R}^{n \times n}$, $B_2 \in \mathbb{R}^{n \times q}$, $E \in \mathbb{R}^{n \times r}$, $F_2 \in \mathbb{R}^{r \times m}$, and $G_2 \in \mathbb{R}^{r \times m}$ are unknown constant matrices with the pair $(AB)$ stabilizable and $G$ nonsingular; $\omega e^{R}$ is the state of the dynamic uncertainty; $\Delta_1 = D_1(w, y)$ and $\Delta_2 = H_2(w, y)$ are the outputs of the dynamic uncertainty, with $D_1 \in \mathbb{R}^{r \times n}$, $D_2 \in \mathbb{R}^{r \times n}$, $\omega e^{R}$ are two unknown locally Lipschitz functions vanishing at the origin. In addition, assume the upper bounds of the norms of $B, D, H, G$ are known.

Letting $u = u_0 + e(t)$ for all $t \in [t_{i-1}, t_i]$, with $u_0$ an initial stabilizing control policy, the following learning strategy is derived [22]:

**Phase-one learning:**

$$x_i^T P_{1,ki}^{-1} x_i = \int_{t_i}^{t_{i+1}} -x_i^T (Q_1 + K_{1,i} R_1 K_{1,i}^T x_i + 2z_i^T + K_{1,i}^T z_i) + 2\gamma_1 + \Delta_1 (x_i^T K_{1,i}^T z_i) + \sum_{j=1}^{n} \gamma_j z_i^T L_{1,ji} z_i \, dt$$

For the matrix $K_{1,i}$ obtained from Phase-one learning, we define $\xi = z + K_{1,i} x_i$.

**Phase-two learning:**

$$\xi_i^T P_2 j_i \xi_i = \int_{t_i}^{t_{i+1}} -\xi_i^T (Q_2 + K_{2,j} R_2 K_{2,j}^T) z_i + 2\Delta_2 + K_{2,j}^T z_i + 2\gamma_2 z_i + 2z_i^T P_0 \xi_i$$

and $K_{2,j}$ such that $A_i - B_i K_{2,j}$ is Hurwitz, with $u_i = -K_{2,j}^T x_i + e_i(t)$, along the solutions of (21), it follows that [23]

$$x_i^T P_3 x_i = \int_{t_i}^{t_{i+1}} -x_i^T (Q_i + K_{i} R_i K_{i}^T x_i + 2z_i^T + K_{i}^T z_i) + 2\gamma_i z_i^T L_{i} z_i \, dt$$

and $K_i$ is Hurwitz, with $u_i = -K_{i}^T x_i + e_i(t)$, along the solutions of (21), it follows that [23]

**Theorem 10** (Jiang and Jiang [23]). For any 1 s i S, suppose $e_i(t)$ satisfies the PE condition such that (24) has a unique solution $P_i^e(t) = (P_i^e(t))'$ and $k_i(t)$, with $Q_i \geq 0, T_i > 0, C_i^T = C_i, k_i(t)$ are unknown functions satisfying (22), $R_i, C_i$ such that $A_i - B_i K_{i}$ is Hurwitz, with $u_i = -K_{i}^T x_i + e_i(t)$, along the solutions of (21), it follows that [23]

$$x_i^T P_3 x_i = \int_{t_i}^{t_{i+1}} -x_i^T (Q_i + K_{i} R_i K_{i}^T x_i + 2z_i^T + K_{i}^T z_i) + 2\gamma_i z_i^T L_{i} z_i \, dt$$

and $K_i$ is Hurwitz, with $u_i = -K_{i}^T x_i + e_i(t)$, along the solutions of (21), it follows that [23]

$$x_i^T P_3 x_i = \int_{t_i}^{t_{i+1}} -x_i^T (Q_i + K_{i} R_i K_{i}^T x_i + 2z_i^T + K_{i}^T z_i) + 2\gamma_i z_i^T L_{i} z_i \, dt$$

and $K_i$ is Hurwitz, with $u_i = -K_{i}^T x_i + e_i(t)$, along the solutions of (21), it follows that [23]

**Theorem 10** (Jiang and Jiang [23]). For any 1 s i S, suppose $e_i(t)$ satisfies the PE condition such that (24) has a unique solution $P_i^e(t) = (P_i^e(t))'$ and $k_i(t)$, with $Q_i \geq 0, T_i > 0, C_i^T = C_i, k_i(t)$ are unknown functions satisfying (22), $R_i, C_i$ such that $A_i - B_i K_{i}$ is Hurwitz, with $u_i = -K_{i}^T x_i + e_i(t)$, along the solutions of (21), it follows that [23]

$$x_i^T P_3 x_i = \int_{t_i}^{t_{i+1}} -x_i^T (Q_i + K_{i} R_i K_{i}^T x_i + 2z_i^T + K_{i}^T z_i) + 2\gamma_i z_i^T L_{i} z_i \, dt$$

and $K_i$ is Hurwitz, with $u_i = -K_{i}^T x_i + e_i(t)$, along the solutions of (21), it follows that [23]

**6. RADP for genuinely nonlinear systems**

In this section, we consider the RADP theory for genuinely nonlinear systems. More remains to be accomplished in this direction. As a starting point, we consider a general single-input nonlinear system

$$\dot{x} = f(x) + g(x) u$$

where $x \in \mathbb{R}^n$ is the system state, $u \in \mathbb{R}$ is the control input, $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ are locally Lipschitz functions. For any initial condition $x_0 \in \mathbb{R}^n$, the cost function associated with (26) is defined as

$$J(x_0; u) = \int_0^\infty (Q(x(t)) + r(t)^2) \, dt, \quad x(0) = x_0$$

where $Q(\cdot)$ is a positive definite function and $r > 0$ is a constant. It is assumed that there exists an admissible control policy $u = u_0(x)$ in the sense that, under this policy, the system (26) is globally asymptotically stable at the origin and the cost (27) is finite. It is well-known that the optimal control policy subject to (27) can be obtained by solving the following HJB equation:

$$0 = \nabla V(x(x)) + Q(x) - \frac{1}{4r} \nabla V(x(x))^2$$

asymptotically stabilizes (6)–(8) at the origin. Moreover, the controller $u = K^* x$ is optimal with respect to the cost (2) when the dynamic uncertainty characterized by the w-system is void.
with the boundary condition \( V(0) = 0 \). Indeed, if the analytical solution \( V^* \) of (28) can be found, the optimal control policy is given by

\[
u^*(x) = -\frac{1}{2R}g(x)^T \nabla V^*(x).
\] (29)

Unfortunately, it is generally impossible to obtain the analytical solution of (28) exactly. However, if \( V^*(x) \) exists, it is always possible to obtain an approximate solution, say, for example, using the policy iteration technique [47], as shown below:

(1) For any integer \( i \geq 0 \) and an admissible policy \( u_i(x) \), solve for \( V_i(x) \) with \( V_i(0) = 0 \) using

\[
0 = \nabla V_i(x)[f(x) + g(x)u_i(x)] + Q(x) + ru_i(x)^2.
\] (30)

(2) Update the control policy using

\[
u_{i+1}(x) = -\frac{1}{2R}g(x)^T \nabla V_i(x).
\] (31)

In the absence of an a priori knowledge of \( f \) and \( g \), the above policy iteration is not implementable. To get around this obstacle, we develop a nonlinear variant of the online policy iteration technique [21].

To begin with, notice that (26) can be rewritten as

\[
x = f(x) + g(x)u(x) + g(x)v_i
\] (32)

where \( v_i = u - u_i \). For each \( i \geq 0 \), the time derivative of \( V_i(x) \) along the solutions of (32) satisfies

\[
V_i(x) = -Q(x) - ru_i^2(x) - 2ru_{i+1}(x)v_i.
\] (33)

Integrating both sides of (33) on any time interval \([t, t+T]\), it follows that

\[
V_i(x(t+T)) - V_i(x(t)) = \int_t^{t+T} [-Q(x) - ru_i^2(x) - 2ru_{i+1}(x)v_i] \, dt.
\] (34)

Notice that if \( u_i(x) \) is given, the unknown functions \( V_i(x) \) and \( u_{i+1}(x) \) can be approximated using (34). To be more specific, for any given compact set \( \Omega \subset \mathbb{R}^n \) containing the origin as an interior point, let \( \{\phi_j(x)\}_{j=1}^\infty \) be an infinite sequence of linearly independent smooth basis functions on \( \Omega \), where \( \phi_j(0) = 0 \) for all \( j = 1, 2, \ldots \). Then, for each \( i = 0, 1, \ldots \), the cost function and the control policy are approximated by \( \hat{V}_i(x) = \sum_{j=1}^{N_i} \hat{c}_{ij}\phi_j(x) \) and \( \hat{u}_{i+1}(x) = \sum_{j=1}^{N_i} \hat{w}_{ij}\phi_j(x) \), respectively, where \( N_i > 0 \) are two sufficiently large integers, and \( \hat{c}_{ij}, \hat{w}_{ij} \) are constant weights to be determined.

Replacing \( V_i(x) \), \( u_i(x) \), and \( u_{i+1}(x) \) in (34) with their approximations, we obtain

\[
\sum_{j=1}^{N_i} \hat{c}_{ij}[\phi_j(x(t_{i+1})) - \phi_j(x(t_i))]
\]

\[
= -\int_{t_i}^{t_{i+1}} 2r \sum_{j=1}^{N_i} \hat{w}_{ij}\phi_j(x) \hat{v}_i \, dt - \int_{t_i}^{t_{i+1}} [Q(x) + ru_i^2(x)] \, dt + e_{i,k},
\] (35)

where \( \hat{v}_i = u_i - \hat{u}_i \), and \( (t_{i+1})_k = 0 \) is a strictly increasing sequence with \( l > 0 \) sufficiently large integer. Then, the weights \( \hat{c}_{ij} \) and \( \hat{w}_{ij} \) can be solved in the minimum of least-squares (i.e., by minimizing \( \sum_{k=1}^{N_i} x_k^2 \)).

Now, starting from \( u_0(x) \), two sequences \( \{\hat{V}_i(x)\}_{i=1}^\infty \), and \( \{\hat{u}_{i+1}(x)\}_{i=1}^\infty \), can be generated via the online policy iteration technique (35). Next, we show the convergence of the sequences to \( V^*(x) \) and \( u^*(x) \), respectively.

**Assumption 11.** There exist \( \omega_0 > 0 \) and \( \delta > 0 \), such that for all \( i \geq 0 \), we have

\[
\frac{1}{T} \sum_{k=0}^{T-1} \theta_{i,k} \geq \delta N_i N_k
\] (36)

where

\[
\theta_{i,k} = \begin{bmatrix} \phi_1(x(t_{k+1})) - \phi_1(x(t_k)) \\ \phi_2(x(t_{k+1})) - \phi_2(x(t_k)) \\ \vdots \\ \phi_N(x(t_{k+1})) - \phi_N(x(t_k)) \\ 2r \int_{t_k}^{t_{k+1}} \phi_1(x) \hat{v}_i \, dt \\ 2r \int_{t_k}^{t_{k+1}} \phi_2(x) \hat{v}_i \, dt \\ \vdots \\ 2r \int_{t_k}^{t_{k+1}} \phi_N(x) \hat{v}_i \, dt \end{bmatrix} \in \mathbb{R}^{N_i N_k}.
\]

**Assumption 12.** For all \( i \geq 0 \), we have \( x(t) \in \Omega \).

Notice that, **Assumption 12** is not very restrictive and can be satisfied if \( \Omega \) is an invariant set for the subsystem.

**Theorem 13** (Jiang and Jiang [27]). Under **Assumptions 11 and 13**, for each \( i \geq 0 \), we have

\[
\lim_{N_i \to \infty} V_i(x) = V^*(x),
\] (37)

\[
\lim_{N_i \to \infty} \hat{u}_{i+1}(x) = u^*(x),
\] (38)

for all \( x \in \Omega \).

Additionally, as a corollary of **Theorem 13**, we can obtain the following convergence properties with respect to the optimal solutions \( V^* \) and \( u^* \).

**Theorem 14** (Jiang and Jiang [27]). Under **Assumptions 11 and 12**, for any arbitrary \( \epsilon > 0 \), there exist positive integers \( i^*, N_{i^*} \) and \( N_{i^*} \), such that

\[
|\hat{V}_i(x) - V^*(x)| < \epsilon \quad \text{and} \quad |\hat{u}_{i+1}(x) - u^*(x)| < \epsilon
\] (39)

for all \( x \in \Omega \), whenever \( i > i^* \), \( N_i > N_{i^*} \) and \( N_k > N_{k^*} \).

It is of interest to note that the main result of this section can also be extended to the case with dynamic uncertainty, which is omitted for want of space.

### 7. Applications

In this section, we demonstrate the wide applicability and the effectiveness of the developed RAPD methodology by means of two practical applications.

#### 7.1. Application to power systems

Consider the classical multimachine power system with governor controllers [35]:

\[
\delta_i(t) = \omega_i(t)
\] (40)

\[
\dot{\omega}_i(t) = -\frac{D_i}{2\tau_i} \omega_i(t) + \frac{w_{0i}}{2\tau_i} [P_m(t) - P_{ai}(t)].
\] (41)

\[
\dot{P}_{mi}(t) = \frac{1}{\tau_i} [-P_{mi}(t) + \omega_i(t)]
\] (42)

\[
P_{ai}(t) = E_{ai} \sum_{j=1}^{N} E_{aj} B_{aj} \sin \delta_j(t) + C_{ai} \cos \delta_i(t)
\] (43)
where \( \delta_i(t) \) is the angle of the \( i \)th generator; \( \delta_j(t) = \delta_i(t) - \delta_j(t) \) is the difference between the angles of the \( i \)th and the \( j \)th generators; \( \omega_0(t) \) is the relative rotor speed; \( P_{m0}(t) \) and \( P_{el}(t) \) are the mechanical power and the electrical power; \( u_0(t) \) is the governor valve opening; \( E_q \) is the transient EMF in quadrature axis, and is assumed to be constant under high-gain SCR controllers; \( D_0, H_0, \) and \( T_i \) are the damping constant, the inertia constant and the governor time constant; \( B_{ij}, C_{ij} \) are constants for \( 1 \leq i,j \leq N \).

Notice that system (40)-(42) can be put into the following form:

\[
\begin{align*}
\Delta \dot{\delta}_i(t) &= \Delta \omega_i(t), \\
\Delta \dot{\omega}_i(t) &= -\frac{D_i}{2H_i} \Delta \omega_i(t) + \frac{\omega_0}{2H_i} \Delta P_m(t), \\
\Delta \dot{P}_m(t) &= \frac{1}{T_i} [\Delta P_m(t) + u_i(t) - d_i(t)]
\end{align*}
\]

where \( \Delta \delta_i(t) = \delta_i(t) - \delta_0, \ \Delta \omega_i(t) = \omega_i(t) - \omega_0, \ \Delta P_m(t) = P_m(t) - P_{m0}(t), \ u_i(t) = u_0(t) - P_{el}(t), \) and

\[
d_i(t) = E_q \sum_{j=1}^{N} B_{ij} \cos \left( \delta_j(t) - \delta_0 \right) \sin \delta_j(t) \times [\Delta \omega_j(t) - \Delta \omega_0(t)].
\]

Therefore, the model (44)-(46) is in the same form as (21) and (22), if we define \( x_i = [\Delta \delta_i(t) \Delta \omega_i(t) \Delta P_m(t)]^T \) and \( y_i = \Delta \omega_0(t) \).

A ten-machine power system is considered for numerical studies [23]. The speed governor controllers and ADP-based learning systems are installed on generators 2–10. In the simulation, generator 1 is used as the reference machine. Parameters of the system are given in [23]. All the parameters, except for the operating points, are assumed to be unknown to the learning systems.

The robust-ADP algorithm is applied to generators 2–10. Trajectories of the angles and frequencies of generators 2–4 are shown in Figs. 3 and 4.

7.2. Application to a single-joint human arm movement control problem

The second application is related to the sensorimotor control problem in computational neuroscience. What we propose here can be considered as a nonlinear extension of the solution presented in [26].

Consider a single-joint arm movement as shown in Fig. 5, where the position of the elbow is fixed. The dynamic model is shown below [48].

\[
i \ddot{\theta} = -mgl \cos(\theta) + n + T_m
\]

where \( m \) is the mass of segment, \( I \) is the inertia, \( g \) is the gravitational constant, \( l \) is the distance of the center of mass from

Fig. 3. Trajectories of the angles.
the joint, $\theta$ is the joint angular position, $T_m$ is the input to the muscle from the motorneurons, and $n$ denotes the inputs from the neural integrator, which can be modeled by a low pass filter as follows with a time constant $\tau_N$.

$$n = -\frac{n}{\tau_N} + T_m.$$  \hspace{1cm} (49)

Let us define $x_1 = \theta - \theta_0$, $x_2 = \dot{\theta}$, $w = n - \tau_N mgl \cos(\theta_0) / (\tau_N + 1) - b x_2$, $u = T_m - mgl \cos(\theta_0) / (\tau_N + 1)$, where $\theta_0$ is the desired end point angular position. Then, the system can be converted to

$$\dot{w} = -\frac{1}{\tau_N} (w + b x_2)$$  \hspace{1cm} (50)

$$-2mgl \sin \left(\frac{\chi_1}{2}\right) \sin \left(\frac{\chi_1}{2} + \theta_0\right)$$  \hspace{1cm} (51)

$$x_1 = x_2$$  \hspace{1cm} (52)

$$x_2 = \frac{2mgl}{I} \sin \left(\frac{\chi_1}{2}\right) \sin \left(\frac{\chi_1}{2} + \theta_0\right) + \frac{1}{I} (u + b x_2 + w)$$  \hspace{1cm} (53)

To apply the proposed RADP method, the basis functions we used are polynomials with degrees less than or equal to five. The invariant set is chosen to contain the region $\{(w, x_1, x_2) : |w| \leq 1, |x_1| \leq 0.8, |x_2| \leq 3.5\}$. Only for simulation purpose, we set $\theta_0 = \pi/4$, $m = 1.65$, $l = 0.179$, $g = 9.81$, $I = 0.0779$. An initial control policy is set to be $u(0) = 1$, $x_1(0) = -\pi/4$, and $x_2(0) = 0$. The optimal cost is specified as $J = \int_0^\infty (100x_1^2 + x_2^2 + u^2) \, dt$.

In this simulation, the convergence is attained after 10 iterations. It can be seen from Fig. 6 that the approximated cost function $V_{10}(x)$ is remarkably reduced compared with the initial approximated cost $V_0(x)$. Also, in Fig. 7, we compare the speed curves under the initial control policy, and the policy after 10 iterations. Clearly, after enough iteration steps, the speed profile becomes a bell-shaped curve which is consistent with experimental observations (see, e.g., [1]).
For nonlinear control systems. Nevertheless, as a computational tool for uncertain complex systems, it is our firm belief that the RAPD theory as advocated here has a wide range of potential applications, such as process control problems in industrial systems [10], networked control systems [68], and biological learning and control [48].

8. Summary and outlook

This paper has presented a brief overview on the state of the art of recent developments in ADP, including especially our initiative and progress on RAPD and its application in power systems and computational neuroscience. Some preliminary results on the RAPD theory for linear and nonlinear systems with dynamic uncertainties have been presented which are new with respect to the past literature of ADP. Under this novel framework of RAPD, computational designs for robust optimal control can be carried out based only on the online data of the state and input variables. Robust global stability and convergence analysis are provided. An extension to large-scale complex systems is also given. The take-home message is that, by taking explicit advantage of techniques from modern nonlinear control, more can be achieved but remains to be accomplished in the development of ADP theory for general nonlinear uncertain systems.

Although the theory of RAPD is motivated somehow by the output ADP problem in which only the outputs are measured, what is presented in this work can only be considered as an intermediate step toward the challenging problem of output ADP. The fundamental obstacle is that learning, observation, and controller design are fully intertwined. Other topics of theoretical importance and great practical interest include the trade-off between performance and robustness and the limitations of RAPD for nonlinear control systems. Nevertheless, as a computational

References