Model-free Approximate Dynamic Programming for Continuous-time Linear Systems

Jae Young Lee, Jin Bae Park, and Yoon Ho Choi

Abstract—In this paper, an online approximate dynamic programming (ADP) technique for completely unknown continuous-time linear systems is proposed to solve the infinite horizon linear quadratic (LQ) optimal control problems. To relax the requirements of the known input coupling matrix, the infinite horizon LQ optimal control problem is converted into the proposed cheap control problem. Then, partially model-free ADP technique given in [8] is employed for this modified problem. In addition, by mathematical analysis, it is shown that the exact solution of proposed cheap control problem is an approximated solution of the LQ optimal control problem. Finally, the simulation results for DC motor are provided to verify the applicability of the proposed ADP algorithm.

Index Terms—approximate dynamic programming, LQR, adaptive critics, adaptive optimal control

I. INTRODUCTION

The researches in approximate dynamic programming (ADP), a family of techniques for solving optimal control problems in forward time, have mostly been focused on discrete-time systems [12]. In discrete-time ADP, since one has suffered the lack of stability, researches on developing stable ADP techniques have been carried out [1], [2], [3], [7], [17], [19], which bring ADP techniques, emerging from the field of computational intelligence, into a part of a control engineering framework. Among the proposed stable ADP techniques, model-free ADP [2], [18], also known as Q-learning, can be considered as an adaptive optimal control method by which the controller converges at last to the optimal controller, which minimizes a prespecified value function. Since adaptive control is in general not optimal, and the optimal controller is designed only in offline fashion, the design of an adaptive optimal controller is a challenging task in the field of control engineering.

Unfortunately, Q-learning is not well-posed in continuous-time domain [15], and thus, several alternative algorithms for model-free learning were proposed [14], [15]. However, the algorithms in [14] and [15] did not guarantee the stability of the closed loop systems, and thus, the proposed algorithms could not be brought into the control tasks. In recent years, Vrabie proposed a new ADP algorithm for the partially unknown continuous time linear systems [8], [9]:

\[ \dot{x} = Ax + Bu, \quad x(0) = x_0, \quad (1) \]

where \( x \in \mathbb{R}^n \) and \( u \in \mathbb{R}^m \) are the state and input vector of the system, respectively; \( x_0 \) is the initial value of \( x(t) \);

II. PRELIMINARIES

A. ADP with Partially Unknown Dynamics

Consider the LQ infinite-horizon optimal control problem:

\[ u^*(t) = \arg \min_{u(t)} V(x_0, u(t)) \]

both \( A \) and \( B \) are matrices with appropriate dimensions, respectively. The proposed algorithm approximates, by iterations, both linear quadratic (LQ) optimal controller and corresponding value function, and was partially model-free in a sense that the resulting iteration can be carried out without internal dynamics (the matrix \( A \)). Furthermore, it showed that i) the proposed algorithm is equal to a Quasi-Newton method, which is the variant of the Kleinman’s algorithm [7], and ii) if the parameters of the approximated value function converges, it must converges to those of the optimal value function. The Kleinman’s iterative algorithm is a Newton method for solving the algebraic Riccati equation (ARE), and is monotonically convergent to the solution of ARE [7]. Since stability is not considered in the early model-free algorithms given in [14] and [15], and the Vrabie’s algorithm given in [8] and [9] is only partially model-free, the real model-free stable ADP for continuous-time systems must be developed.

In this paper, we propose an online model-free ADP technique to solve in forward time the infinite horizon LQ optimal control problems for continuous-time linear systems. First, by using precompensation [13], the traditional LQ optimal control problem is transformed into the proposed cheap control problem, the solution of which approximates the solution of the original LQ optimal control problem. Then, the Vrabie’s ADP method is employed to solve this cheap control problem in forward time. If we assume that the proposed model-free ADP method converges, then, as the Vrabie’s ADP algorithm does, it must converge to the optimal solution of the cheap control problem. In the proposed cheap control problem, a dynamic controller is proposed with a new control input, and a quadratic value function for the cheap control problem is formulated based on the value function given in the primary LQ optimal control problem. In addition, by the mathematical analysis, it is shown that the solution, state trajectories, and corresponding value function of the proposed cheap control problem approximate those of the primary LQ optimal control problem.
subject to the dynamics (1), where the value function $V(x_0, u(\cdot))$ is defined by

$$V(x_0, u(\cdot)) = \int_{t_0}^{\infty} r(x, u) \, dt.$$  \hfill (2)

Here, $r(x, u) = x^T Q x + u^T R u$ with $Q \succeq 0$ and $R > 0$. By the Bellman’s optimality principle, the optimal control law $u^*(t)$ and the corresponding optimal value function $V^*(x_0)$ are determined as

- **Optimal value function**: $V^*(x) = x^T P x$,  \hfill (3)
- **Optimal control law**: $u^*(t) = -R^{-1} B^T P x(t)$,  \hfill (4)

where $P \succeq 0$ is the unique solution of the ARE:

$$A^T P + PA - PBR^{-1} B^T P + Q = 0. \hfill (5)$$

The existence of the unique positive definite (semidefinite) solution $P$ is guaranteed if the pair $(A, B, Q^{1/2})$ is stabilizable and observable (detectable) [5]. Applying $u^*(t)$ to the system (1), we obtain the closed loop system:

$$\dot{x} = (A - BR^{-1} B^T P)x.$$  \hfill (6)

The objective of the ADP-based agent given in [9] is to find the above optimal control law $u^*(t)$ online without any knowledge about the system matrix $A$. The value function $V(x(t))$ for a given policy can be approximated as

$$V(x(t)) = \int_{t_0}^{t+T} r(x, u) \, dt + W(x(t+T)), \hfill (7)$$

where $W(x(t+T))$ is an approximation of the value function $V(x(t+T))$ from $t+T$ to $\infty$. Based on (7), the ADP iteration scheme given in [8] and [9] can be derived as follows:

$$V_{i+1}(x(t)) = \int_{t}^{t+T} r(x, u_i) \, dt + V_i(x(t+T)), \quad u_i(t) = -R^{-1} B^T P_i x(t), \hfill (8)$$

where $i$ indicates the iteration number, $u_i(t)$ is the input at the $i$-th iteration, and $V_i(x(t))$ is the approximated value function at the $i$-th iteration, respectively. The information about the system matrix $A$ is not required to implement the above iteration algorithm (the knowledge about the input coupling matrix $B$ is needed). By parameterizing the approximated value function as $V_i(x(t)) = x^T(t) P_i x(t)$, (8) can be rewritten as

$$x^T(t) P_{i+1} x(t) = \int_{t}^{t+T} r(x, u_i) \, dt + x^T(t+T) P_i x(t+T). \hfill (9)$$

Based on (9), the matrix $P_i$ is updated to the matrix $P_{i+1}$ by using a least squares method at each $i$-th iteration step with a sufficient number of points in the state and input trajectories. Vrabie [8] showed that i) the above iteration is equivalent to a Quasi-Newton method, and ii) if $P_i$ converges to $P^*$, $P^*$ satisfies the ARE equation (5) with $P^* = P$. For more details, see [8] and [9].

**Remark 1**: In comparison with [14] and [18], the above algorithm is proved to have convergence properties when it converges, but the $B$ matrix must be exactly known for iterating (8) and (9).

### B. Singular Perturbation Theory

In this subsection, we briefly discuss the singular perturbation theory for autonomous systems, which is employed for developing the main theorem of this paper. Consider the singular perturbation problem of the following autonomous singular perturbation model:

$$\dot{x} = f(x, z, \varepsilon), x(0) = x_0, \hfill (10)$$

$$\varepsilon \dot{z} = g(x, z, \varepsilon), z(0) = z_0 \hfill (11)$$

where $f : D_x \times D_z \times \mathbb{R} \to \mathbb{R}^n$ and $g : D_x \times D_z \times \mathbb{R} \to \mathbb{R}^m$ are vector-valued continuously differentiable functions; $D_x \subset \mathbb{R}^n$ and $D_z \subset \mathbb{R}^m$ are open connected sets. Assume that the solution of $0 = g(x, z, 0)$ has $k \geq 1$ isolated roots. Let $\bar{x}(x)$ be any solution of $0 = g(x, z, 0)$. Then, the reduced model and its boundary-layer model are represented as [10]

**Reduced model:**

$$\dot{\bar{x}} = f(\bar{x}, z(\bar{x}), 0), \quad \bar{x}(0) = x_0 \hfill (12)$$

**Boundary-layer model for fixed $x$:**

$$\frac{dy}{dt} = g(x, y, z(x), 0), \quad y(0) = z_0 - \bar{z}(x_0). \hfill (13)$$

where $\bar{x} \in \mathbb{R}^n$, $y := z - \bar{z}(x)$, and $\tau := t/\varepsilon$, respectively. We define uniform exponential stability in frozen parameters and state a singular perturbation lemma for the system (10) and (11).

**Definition 1**: The equilibrium point $y = 0$ of the boundary-layer system (13) is exponentially stable, uniformly in $x \in D_x$, if $\exists$ positive constants $k, \gamma, \rho_0 > 0$ such that the solutions of (13) satisfy

$$\|y(\tau)\| \leq k \|y(0)\| \exp(-\gamma \tau),$$

$\forall \|y(0)\| \leq \rho_0, \forall x \in D_x, \forall \tau \geq 0$.

**Lemma 1**: Consider the singular perturbation problem of (10) and (11). Assume that the following conditions are all satisfied $\forall [x, z - \bar{z}(x), \varepsilon] \in D_x \times D_z \times [0, \varepsilon_0]$, for some domains $D_x \subset \mathbb{R}^n$ and $D_z \subset \mathbb{R}^m$ containing origins:

- on any compact subset of $D_x \times D_z$, the functions $f$, $g$, and their first partial derivatives with respect to $(x, z, \varepsilon)$ are continuous and bounded, $\bar{z}(x)$ and $\partial g(x, z, 0)/\partial z$ have bounded first partial derivatives with respect to their arguments, and $\partial f(x, z(x), 0)/\partial x$ is Lipschitz.
- the origin is an exponentially stable equilibrium point of the reduced system (12); there is a Lyapunov function $L(x)$ and a positive definite function $W(x)$ such that

$$\partial L/\partial x f(x, z(x), 0) \leq -W(x) \hfill (14)$$

is satisfied $\forall x \in D_x$, and \{ $L(x) \leq c$ \} is a compact subset of $D_x$ for some $c > 0$.
- the origin is an exponentially stable equilibrium point of the boundary-layer system (13), uniformly in $x \in D_x$; let $\mathcal{R}_y \subset \mathcal{R}_y$ be the region of attraction of (11) and $\Omega_y$ be a compact subset of $\mathcal{R}_y$.

Then, for each compact set $\Omega_y \subset \{ L(x) \leq \rho c, \ 0 < \rho < 1 \}$, there is a positive constant $\varepsilon^*$ such that $\forall x_0 \in \Omega_y, z_0 -
that whenever $\varepsilon \in (0, \varepsilon^*)$, the equation (10) and (11) has a unique solution $x(t, \varepsilon)$ and $z(t, \varepsilon)$, and

$$x(t, \varepsilon) - \bar{x}(t) = O(\varepsilon)$$

holds. Moreover, given any $t_0 > 0$, there is $\varepsilon^{**} \leq \varepsilon^*$ such that whenever $\varepsilon < \varepsilon^{**}$,

$$z(t, \varepsilon) - \bar{z}(\bar{x}(t)) = O(\varepsilon)$$

holds, uniformly in $t \in [t_0, \infty)$.

Proof: This is the straight-forward simplification of the original singular perturbation theory on the infinite time interval [10]. So, the proof is omitted. For the complete proof, see [10].

III. CONTINUOUS-TIME MODEL-FREE ADP

In this section, we proposed a model-free ADP technique to solve the infinite-horizon LQ optimal control problem. The iteration scheme (8)-(9) can be extended to the LTI system with unknown matrices $A$ and $B$. Suppose the control input $u$ is represented by the following dynamic equation:

$$\dot{u} = Fx + Gu + v, \quad u(0) = u_0$$

(15)

where $F \in \mathbb{R}^{m \times m}$ and $G \in \mathbb{R}^{m \times m}$ are the suitably chosen matrices, $v \in \mathbb{R}^m$ is the new control input which drives $u$, and $u_0$ is the initial condition for (15), respectively. Now, (15) can be combined with the original plant, and the resulting augmented system can be represented as

$$\dot{z} = \bar{A}z + \hat{B}v := \begin{bmatrix} A & B \\ F & G \end{bmatrix} z + \begin{bmatrix} 0 \\ I_m \end{bmatrix} v,$$

(16)

with the initial condition

$$z(0) = z_0 := \begin{bmatrix} x_0^T \\ u_0^T \end{bmatrix}^T,$$

where $z := \begin{bmatrix} x^T \\ u^T \end{bmatrix}^T \in \mathbb{R}^{(m+n)}$ is the augmented state, $I_m$ is the $m$ by $m$ identity matrix, and $z_0 \in \mathbb{R}^{(m+n)}$ is the initial value of $z(t)$, respectively. Note that all the coefficients of the original plant are merged into $\bar{A}$. Due to the new input $v$, the original value function (2) is modified into the following augmented value function $\bar{V}$:

$$\bar{V}(z_0, v(\cdot)) := \int_0^{\infty} \bar{r}(z, v) \, d\tau$$

(17)

where $\bar{r}(z, v) := z^T \hat{Q}z + \varepsilon^2 v^T Rv, \hat{Q} := \text{blockdiag}[Q, R] \in \mathbb{R}^{(n+m) \times (n+m)}$, and $\varepsilon > 0$, respectively. We introduce $\varepsilon$ to regulate the weight of the new control input $v$. Now, to derive a model-free ADP algorithm, define the cheap control problem:

$$v^*(t) = \arg\min_{v(t)} \bar{V}(z_0, v(\cdot))$$

(18)

subject to the dynamics (16). In Section IV, we will show the relationships of the the solutions, state trajectories, and corresponding value functions for two optimal control problems (2) and (18) for sufficiently small $\varepsilon$. As in Section II, the optimal value function $V^*(z) := \min_u V(z, v(\cdot))$ and the optimal control law $u^*(x, u)$ for this proposed cheap control problem are determined by

Optimal value function: $\bar{V}^*(z) = z^T \bar{P}z$

(19)

Optimal control law:

$$v^*(t) = -\varepsilon^{-2} R^{-1} (\bar{P}_{11}^T x(t) + \bar{P}_{22} u(t))$$

(20)

where $\bar{P} := \begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{12}^T & \bar{P}_{22} \end{bmatrix} \succeq 0$ is the solution of the ARE:

$$\varepsilon^2 (A^T \bar{P} + \bar{P} A + \bar{Q}) = \begin{bmatrix} \bar{P}_{12} R^{-1} \bar{P}_{12}^T & \bar{P}_{12} R^{-1} \bar{P}_{22} \\ \bar{P}_{22} R^{-1} \bar{P}_{12}^T & \bar{P}_{22} R^{-1} \bar{P}_{22} \end{bmatrix}.$$
for the $j$-th element $z_j$ of $z(t)$, and the $kl$-th element $p_{kl}(t)$ of $P$ where $j, k, l = 1, 2, \ldots, n + m$, respectively. By (25) and (26), (24) can be rewritten as

$$
\tilde{z}^T(t)\tilde{p}_{i+1} = d(z(t), P_i).
$$

If the states $\tilde{z}(t)$ and the values of $d(\tilde{z}(t), P_i)$ are obtained at $N (\geq N_{min})$ points in the trajectory, the least-squares problem can be solved by the following equation:

$$
\tilde{p}_{i+1} = (ZZ^T)^{-1}ZY
$$

where

$$Z := \{\tilde{z}^{(1)}, \tilde{z}^{(2)}, \ldots, \tilde{z}^{(N)}\}$$

$$Y := \{d(\tilde{z}^{(1)}, P_i), d(\tilde{z}^{(2)}, P_i), \ldots, d(\tilde{z}^{(N)}, P_i)\}.$$

The requirement $N \geq N_{min}$ is the necessary condition for the excitation condition, namely, the existence of $(ZZ^T)^{-1}$ where $N_{min}$ is the dimension of $P_i$. Excitation condition is needed when we calculate the least-squares solution (27) to obtain $P_{i+1}$ at $i$-th iteration. In the real applications, both $\tilde{z}(t)$ and $d(\tilde{z}(t), P_i)$ are obtained at every $T$ seconds, and thus the least-squares solver determines $P_{i+1}$ at least after $N_{min}T$ seconds are elapsed.

In Fig. 1, the above dynamic equation $\dot{V} = \tilde{z}^TQ \tilde{z} + \varepsilon_2 \dot{z}^T R \dot{z}$ is implemented in the critic, and the values $u, \dot{V}$, and $x$ are sampled at every $T$ seconds in order to solve the least-squares problem (27). After (27) is solved, the updated parameters $P_i$ then flow to the actor with zero order hold (ZOH), and the actor uses $\tilde{P}_i$ in evaluating the new control input $v = v_i$. Then, the new control input $v$ again generates the actual control input $u$ by the law $\dot{u} = Fx + Gu + v$.

Remark 2: As in discrete-time Q-learning case [2], the proposed continuous-time model-free scheme updates $(n + m)(n + m + 1)/2$ parameters in one iteration.

Remark 3: Since the new control input $v$ is not considered in the LQ optimal control problem (2), it is reasonable to choose the value function (17) with the cheap control context [4, 16] such that the new control input $v$ is arbitrarily cheap, i.e., $\varepsilon$ is arbitrarily small. Since $n > m$, $\lim_{\varepsilon \to 0} P$ exists [11], and thus, $\varepsilon$ may be taken to be arbitrarily small. Intuitively, the smaller $\varepsilon > 0$, the smaller difference $V^*(z(t))-V^*(x(t))$ is expected from (17). However, when $A - B \dot{R}^{-1}B^T \dot{P}_0$ is unstable as well as $\varepsilon \approx 0$, $z$ may diverge very rapidly since the control input $v$ becomes extremely large as $\varepsilon$ goes to 0. This deteriorates the initial performance and even cause instability of the system. Therefore, when one selects small $\varepsilon > 0$, care must be taken to prevent $z$ from diverging to infinity in very short time.

Remark 4: In [6], it is stated that the dynamic state feedback may improve the gain and phase margin of the whole closed loop systems, and thus, gives robustness to the closed loop systems. Since the optimal dynamic control law $\dot{u} = (F - \varepsilon_2 \dot{R}_1^{-1} \dot{P}_{12})u + (G - \varepsilon_2 \dot{R}_1^{-1} \dot{P}_{22})u$ is indeed expressed in a dynamic state feedback form $u = S_{11}u + S_{12}z$, $\dot{z} = S_{21}z + S_{22}x$, where $z \in \mathbb{R}^n$, $S_{11} = I_m$, $S_{12} = 0$, $S_{21} = G - \varepsilon_2 \dot{R}_1^{-1} \dot{P}_{12}$, and $S_{22} = F - \varepsilon_2 \dot{R}_1^{-1} \dot{P}_{12}$, it may also improve its robustness properties.

IV. ANALYSIS OF THE CHEAP CONTROL PROBLEM

In this section, we analyze the solution, state trajectories, and optimal value function for the cheap control problem (18) defined in the previous section. As a result, when $\varepsilon$ is sufficiently small, it is proved that the two optimal control problems (2) and (18) are approximately same. Consider the augmented LTI system with the optimal input $v^* = -\varepsilon^{-2}R^{-1}(\tilde{P}_{12}^T x + \tilde{P}_{22} u)$:

$$
\begin{bmatrix}
\dot{x} \\
\dot{u}
\end{bmatrix} =
\begin{bmatrix}
A & B \\
\varepsilon F - R^{-1} \tilde{P}_{12}^T & \varepsilon G - R^{-1} \tilde{P}_{22}
\end{bmatrix}
\begin{bmatrix}
x \\
u
\end{bmatrix}.
$$

Here, $\tilde{P}_{12}$ and $\tilde{P}_{22}$ are the submatrices of $\tilde{P}(\varepsilon)$; $\tilde{P}(\varepsilon)$ is defined for sufficiently small $\varepsilon$ by

$$
\tilde{P}(\varepsilon) := \begin{bmatrix}
\tilde{P}_{11} & \varepsilon \tilde{P}_{12} \\
\varepsilon \tilde{P}_{12} & \tilde{P}_{22}
\end{bmatrix} = \begin{bmatrix}
\tilde{P}_{11} & \tilde{P}_{12} \\
\tilde{P}_{12} & \tilde{P}_{22}
\end{bmatrix} = \tilde{P}(\varepsilon).
$$

where $\tilde{P}_{11} := \tilde{P}_{11}, \tilde{P}_{12} := \varepsilon^{-1} \tilde{P}_{12}$, and $\tilde{P}_{22} := \varepsilon^{-1} \tilde{P}_{22}$, respectively. We now analyze the above system for sufficiently small $\varepsilon$ to find the relationships between two closed-loop systems (6) and (28).

Lemma 2: Assume that $(A, B, Q^{1/2})$ is stabilizable and detectable. Then, for sufficiently small $\varepsilon > 0$, the submatrices $\tilde{P}_{11}, \tilde{P}_{12},$ and $\tilde{P}_{22}$ satisfies the followings:

\begin{align*}
i) & \quad \tilde{P}_{11}(\varepsilon) = P + O(\varepsilon) \quad (30) \\
ii) & \quad \tilde{P}_{12}(\varepsilon) = PB + O(\varepsilon) \quad (31) \\
iii) & \quad \tilde{P}_{22}(\varepsilon) = R + O(\varepsilon). \quad (32)
\end{align*}

Proof: By using (29), the ARE equation (21) can be rearranged as

$$
\begin{align*}
\tilde{P}_{11} A + A^T \tilde{P}_{11} + \varepsilon \tilde{P}_{12} F + \varepsilon F^T \tilde{P}_{12} + Q &= \tilde{P}_{12} R^{-1} \tilde{P}_{22}^T \quad (33) \\
\tilde{P}_{11} B + \varepsilon \tilde{P}_{12} G + \varepsilon A^T \tilde{P}_{12} + \varepsilon F^T \tilde{P}_{22} &= \tilde{P}_{12} R^{-1} \tilde{P}_{22} \quad (34) \\
\varepsilon \tilde{P}_{12} B + \varepsilon B^T \tilde{P}_{12} + \varepsilon \tilde{P}_{22} G + \varepsilon G^T \tilde{P}_{22} + R &= \tilde{P}_{22} R^{-1} \tilde{P}_{22}. \quad (35)
\end{align*}
$$
For sufficiently small $\varepsilon$, each $P_{ij}$ can be represented by [4]

\begin{align}
\dot{P}_{11}(\varepsilon) &= \dot{P}_{110} + O(\varepsilon) \\
\dot{P}_{12}(\varepsilon) &= \dot{P}_{120} + O(\varepsilon) \\
\dot{P}_{22}(\varepsilon) &= \dot{P}_{220} + O(\varepsilon)
\end{align}

(36) (37) (38)

where $\dot{P}_{ij0}$, $(i, j = 1, 2)$ are the constant matrices to be determined. Let $\varepsilon$ be zero to calculate $\dot{P}_{ij0}$. Then, from (33)~(35), we have

\begin{align}
\dot{P}_{110}A + A^T\dot{P}_{110} + Q &= P_{120}R^{-1}\dot{P}_{120}^T \\
\dot{P}_{110}B &= \dot{P}_{120}R^{-1}\dot{P}_{220} \\
R &= \dot{P}_{220}R^{-1}\dot{P}_{220}.
\end{align}

(39) (40) (41)

From (40) and (41), we obtain $\dot{P}_{220} = R$ and $\dot{P}_{120} = \dot{P}_{110}B$. Thus, substituting $\dot{P}_{120} = \dot{P}_{110}B$ into (39) yields

\[ A^T\dot{P}_{110} + \dot{P}_{110}A - \dot{P}_{110}BR^{-1}B^T\dot{P}_{110} + Q = 0, \]

which is the same form as the original ARE (5). Since we assume that $(A, B)$ is stabilizable and $(Q^{1/2}, A)$ is detectable, (5) has a unique positive semidefinite solution. Therefore, we conclude that $\dot{P}_{110} = P$, $\dot{P}_{120} = PB$, and $\dot{P}_{220} = R$, which completes the proof.

Lemma 2 shows the first terms in the Taylor expansions of the submatrices $\dot{P}_{11}$, $\dot{P}_{12}$, and $\dot{P}_{22}$ with respect to $\varepsilon$ is. The following lemma shows that if $\varepsilon$ goes to zero, the states of the closed control problem converges to the states of the primary LQ optimal control and the corresponding control law.

**Lemma 3:** Assume that $(A, B)$ is stabilizable and $(Q^{1/2}, A)$ is detectable. If $\varepsilon$ goes to zero, the state $[x, u]^T$ converges to $[\bar{x}, \bar{u}]^T$ where $\bar{x}$ and $\bar{u}$ are defined as follows:

\begin{align}
\dot{\bar{x}} &= (A - BR^{-1}B^TP)\bar{x} \\
\dot{\bar{u}} &= u*(\bar{x}) = -R^{-1}B^TP\bar{x}. 
\end{align}

(42) (43)

**Proof:** Letting $\varepsilon \to 0$ in (28) yields

\begin{align}
\dot{x} &= A\bar{x} + B\bar{u} \\
0 &= -R^{-1}(\dot{\bar{P}}_{12}^T\bar{x} + \dot{\bar{P}}_{22}\bar{u}).
\end{align}

(44) (45)

When $\varepsilon \to 0$, $\dot{P}_{12} = PB$ and $\dot{P}_{22} = R$ are satisfied by Lemma 2. Substituting these into (45) and rearranging the equation (44) and (45) yield (42) and (43).

In Lemma 3, (42) and $\bar{u}$ are called the reduced model and corresponding quasi-steady-state, respectively. Now, here is the main theorem of this paper revealing the relationships between two optimal control problems (2) and (18).

**Theorem 1:** Consider the system (28). Assume that $(A, B)$ is stabilizable and $(Q^{1/2}, A)$ is observable. Let $x(t, \varepsilon)$ and $u(t, \varepsilon)$ be the unique solution of the system (28). Then, there exists a constant $\varepsilon^* > 0$ such that for all $t \in [0, \infty)$ and $0 < \varepsilon < \varepsilon^*$,

\begin{align}
i) & \quad \bar{V}*(x, u, \varepsilon) - V*(x) = O(\varepsilon) \\
ii) & \quad x(t, \varepsilon) - \bar{x}(t) = O(\varepsilon)
\end{align}

(46) (47)

hold. Furthermore, for all $\bar{t} > 0$, there exists $\varepsilon^* < \varepsilon^*$ such that for all $t \in [\bar{t}, \infty)$ and $0 < \varepsilon < \varepsilon^*$,

\[ u(t, \varepsilon) = u^*(t) = O(\varepsilon) \]

(48)

holds.

**Proof:** Note that by Lemma 2, $\dot{P}$ is represented as

\[ \dot{P} = \begin{bmatrix} P + O(\varepsilon) & \varepsilon PB + O(\varepsilon^2) \\
\varepsilon B^TP + O(\varepsilon^2) & \varepsilon R + O(\varepsilon^2) \end{bmatrix} = \begin{bmatrix} P & 0 \\
0 & 0 \end{bmatrix} + O(\varepsilon), \]

for sufficiently small $\varepsilon$. Then, the optimal value function (19) for the system (16) is directly computed as

\[ \bar{V}*(z, \varepsilon) = z^TPz = x^TPx + O(\varepsilon), \]

which implies $V*(x, \varepsilon) - V*(x) = O(\varepsilon)$, the first part of this theorem.

Now, we will apply Lemma 1 to develop the remaining part of this theorem. In this case, the functions $f$ and $g$ in Lemma 1 is determined by

\[ f(x, u) = Ax + Bu, \]
\[ g(x, u, \varepsilon) = \varepsilon(Fx + Gu) - R^{-1}(\bar{P}_{12}^T\varepsilon)x + \bar{P}_{22}\varepsilon), \]

where $f$ and $g$ satisfies the followings:

\begin{align}
\frac{\partial f(x, u)}{\partial x} &= A, \quad \frac{\partial f(x, u)}{\partial u} = B, \\
\frac{\partial g(x, u, \varepsilon)}{\partial x} &= A - BR^{-1}B^TP, \\
\frac{\partial g(x, u, \varepsilon)}{\partial u} &= \varepsilon F - R^{-1}\bar{P}_{12}(\varepsilon), \\
\frac{\partial g(x, u, \varepsilon)}{\partial x} &= \varepsilon G - R^{-1}\bar{P}_{22}(\varepsilon), \\
\frac{\partial g(x, u, \varepsilon)}{\partial \varepsilon} &= (F - R^{-1}\bar{P}_{12}(\varepsilon))x + (G - R^{-1}\bar{P}_{22}(\varepsilon))u, \\
\frac{\partial g(x, u, 0)}{\partial \varepsilon} &= -I.
\end{align}

(49) (50) (51) (52) (53) (54)

Here note that the partial derivatives of $\bar{P}_{12}(\varepsilon)$ and $\bar{P}_{22}(\varepsilon)$ in (53) exist for sufficiently small $\varepsilon$, since the matrix $\dot{P}(\varepsilon)$ has Taylor series expansion when $\varepsilon$ is sufficiently small [4][16]. Now, define the small constant $\varepsilon_1 > 0$ such that for all $\varepsilon \in [0, \varepsilon_1], \dot{P}(\varepsilon)$ have Taylor series expansion. Note that Taylor expansion of $\dot{P}(\varepsilon)$ contains (36) through (38) as a special case. This means that for any $\varepsilon \in [0, \varepsilon_1]$, Lemma 2 holds, and thus, for any $\varepsilon \in [0, \varepsilon_1]$, the first part of this theorem, i.e., $V*(x, u, \varepsilon) - V*(x) = O(\varepsilon)$ also holds. Now, define the variable $e_u$ as

\[ e_u = u - \bar{u}(x) \]
\[ = u - u^*(x) = u + R^{-1}B^TPx. \]

And also define the set $S \subset \mathbb{R}^{n+m+1}$ as

\[ S = \{ s \in \mathbb{R}^{n+m+1} | s \in S_x \times S_{e_u} \times [0, \varepsilon_1] \}. \]
where $S_\varepsilon$ and $S_{\varepsilon_0}$ are any compact subset of $\mathbb{R}^n$ and $\mathbb{R}^m$ containing both initial conditions $x_0$ and $u_0 - u(x_0)$ as well as their respective origins. Then, for all $s := (x, e_u, \varepsilon) \in S$, the followings are true:

- the functions $f(x, u)$ and $g(x, u, \varepsilon)$, and their partial derivatives (49), (51)-(53) are continuous and bounded;
- $\bar{u}(x) = -R^{-1}B^TPx$ and (54) have bounded first partial derivatives with respect to their arguments;
- $\partial f(x, u(x))$$/\partial x$ in (50) is obviously Lipschitz in $x$.

Therefore, the first part of Lemma 1 is satisfied.

Now, we proceed with guaranteeing the second part of Lemma 1. By Lemma 3, the reduced model and the quasi-steady-state are given by (42) and (43), respectively. Since we assume that $(A, B, Q^{1/2})$ is stabilizable and observable, (42) is globally exponentially stable and a radially unbounded Lyapunov function $L(x)$ for the reduced model (42) is given by $L(x) = V^*(x) = x^TPx$ [5]. Since $L(x)$ is radially unbounded, $\{L(x) \leq c\} \subset S_x$ is compact for all $c > 0$. Therefore, the global exponential stability of (42) and the existence of its Lyapunov function imply that the second condition of Lemma 1 is satisfied for any $c > 0$.

Lastly, we analyze the boundary layer system to examine whether the third condition is met. The boundary layer system in the $\tau$ time scale can be obtained as [10]

$$\frac{de_u}{d\tau} = g(x, e_u + \bar{u}, 0) = -R^{-1}(B^TPx + R(e_u + \bar{u})) = -e_u,$$

which is obviously globally exponentially stable, uniformly in $x \in \mathbb{R}^n$. Therefore, the last condition in Lemma 1 is also satisfied with the region of attraction $\mathcal{R}_g = \mathbb{R}^m$. Now, choose $c > 0$ such that the compact set $\{L(x) \leq c\}$ contains the initial condition $x_0$. For any $x_0 \in \mathbb{R}^n$, the compact subset $S_x \subset \mathbb{R}^n$ and the constant $c > 0$ can always be chosen such that the compact set $\{L(x) \leq c\} \subset S_x$ contains $x_0$ as well as the origin. In the same way, for any $u_0 - u(x_0) \in \mathbb{R}^m$, the compact subset $S_{\varepsilon_0}$ can always be chosen such that $S_{\varepsilon_0}$ contains the initial condition $u_0 - u(x_0)$ as well as the origin. Since all the conditions in Lemma 1 are satisfied for any compact subsets $\{L(x) \leq c\} \subset S_x$ and $S_{\varepsilon_0} \subset \mathbb{R}^m$ defined in the above, there exists a positive constant $\varepsilon_0^*$ such that i) for all $\varepsilon \in (0, \varepsilon_0^*)$, $x(t, e) - x(t) = O(\varepsilon)$ is satisfied, and ii) given any $t_b > 0$, there is $\varepsilon_0^{**} \leq \varepsilon_0^*$ such that $\forall t \in [t_b, \infty)$ and $\forall \varepsilon \in (0, \varepsilon_0^{**})$, $u(t, e) - u^{*}(t) = O(e)$ holds.

Since $\dot{V}^*(x, u, \varepsilon) - V^*(x) = O(\varepsilon)$ holds $\forall \varepsilon \in (0, \varepsilon_0^*)$, the proof is completed with $e^* = \min\{\varepsilon_0^*, \varepsilon_0^{**}\}$ and $e^* = \min\{\varepsilon^*, e^*\}$.

V. SIMULATION RESULTS

We apply the proposed ADP scheme to the idealized linear DC motor model:

$$\dot{x} = \begin{bmatrix} -b/J & K/J \\ -K/L & -R/L \end{bmatrix}x + \begin{bmatrix} 0 \\ 1/L \end{bmatrix}u$$

with the associated value function

$$V(x) := \int_0^\infty x_1^2 + x_2^2 + 0.5u^2 \, d\tau.$$

where $K$ is the motor constant, $L$ and $R$ are the electric inductance and the resistance in the DC motor, $b$ is the motor friction constant, and $x := [\theta, i]^T$; $\theta$ and $i$ are the motor angle and the electric current, respectively. By appropriate selection of DC motor parameters, the system dynamic matrices are determined as

$$A = \begin{bmatrix} -5 & 1 \\ -0.02 & -20 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 20 \end{bmatrix}. \quad (56)$$

And, the corresponding Ricatti matrix $P$ and LQ optimal controller $u$ are calculated as

$$P = \begin{bmatrix} 0.0995 & 0.0025 \\ 0.0025 & 0.0184 \end{bmatrix}, \quad u = -0.0999 x_1 - 0.7349 x_2. \quad (57)$$

First, we examine the approximation capability of the cheap controller. The DC motor is controlled by the dynamic form:

$$\dot{u} = -\theta - u + v, \quad (58)$$

with which the augmented dynamic system is represented as

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{u} \end{bmatrix} = \begin{bmatrix} -5 & 1 & 0 \\ -0.02 & -20 & 20 \\ -1 & 0 & -1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ u \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} v. \quad (59)$$

By choosing $\varepsilon^2 = 0.02$, The value function for the cheap control problem (18) is formulated as

$$\bar{V}(x, u) := \int_0^\infty x_1^2 + x_2^2 + 0.5u^2 + 0.01v^2 \, d\tau. \quad (60)$$

By combining (59) with (60), the Ricatti matrix $\bar{P}$ and the optimal control input $v^*$ for this problem is determined by

$$\bar{P} := \begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{12} & \bar{P}_{22} \end{bmatrix} = \begin{bmatrix} 0.1003 & 0.0035 & -0.0018 \\ 0.0035 & 0.0245 & 0.0159 \\ -0.0018 & 0.0159 & 0.0971 \end{bmatrix}$$

$$v = 0.1784 x_1 - 1.5924 x_2 - 9.7096 u. \quad (59)$$

Note that $\bar{P}_{11}$ is approximately equal to $P$. This is because for sufficiently small $\varepsilon > 0$, $\bar{P}_{11} \approx P$ holds according to Lemma 2.

Fig. 2 demonstrates the state trajectories of the two optimal controller when the initial conditions are set to $x(0) = [5, 2]^T$ and $u(0) = 0$. In Fig. 2, both trajectories for state $x_1$ are almost same, and there is a small delay for $x_2$ when the proposed cheap controller is applied. The trajectories of the control inputs for both cases are described in Fig. 3. As seen in Fig. 3-(a), the input $u$ for the cheap case cannot be changed rapidly since the dynamic feedback (58) acts as a low pass filter. The smaller $\varepsilon > 0$, the more rapid response in the control input $u$ can be obtained.

We now apply the proposed online ADP to the augmented system (59) so as to find the optimal gain matrix $P$. Note that in this case, any knowledge about the matrices in the DC motor dynamics is not required to perform the iteration given by (22) and (23). In this ADP iteration, the initial gain matrix $\bar{P}_0$ is set to $\bar{P}_0 = 0$.

The simulation results are illustrated in Figs. 4-7. In this simulation, the period $T$ is set to 25 ms, and thus, the
least-squares method is employed at every 300 ms. For each iteration, $6 (= N_{\text{min}})$ points are collected to determine $P_t$. Fig. 5 shows the evolution of the critic parameters. The final parameter value obtained by the proposed ADP is 

$$\tilde{P}_f = \begin{bmatrix} 0.1007 & 0.0035 & -0.0018 \\ 0.0035 & 0.0246 & 0.0160 \\ -0.0018 & 0.0160 & 0.0974 \end{bmatrix}.$$  

The maximum difference of each element of the matrices $\tilde{P}$ and $P_f$ is $4 \times 10^{-4}$, which shows the convergence of the proposed algorithm. The state trajectories and control inputs $u$ and $v$ are given in Figs. 4, 6, and 7. In those figures, the marked points indicate that the iterative update (22) and (23) is executed at that time. Note that we obtain a suboptimal controller without any predesigned initial controller and identification process in this simulation.

VI. CONCLUSIONS

This paper presented an online model-free ADP technique to solve an infinite horizon LQ optimal control problem for unknown continuous-time linear systems. The proposed ADP had the actor-critic structure as in the Vrabie’s ADP, but instead of the traditional LQ optimal control problem, the proposed cheap control problem defined with a dynamic controller and a modified value function was solved without requiring any information about the matrices in the system dynamics. By mathematical analysis, it was verified that the approximated solution for the LQ optimal control problem can be obtained by solving the proposed cheap control problem if $\varepsilon$ is sufficiently small. The simulation for an ideal DC motor was presented to verify the applicability of the proposed model-free ADP method.

REFERENCES

Fig. 4. The trajectories of the states $x_1$ and $x_2$ when the proposed model-free ADP technique is applied.

Fig. 5. The critic parameter variations when the proposed model-free ADP technique is applied.

Fig. 6. The trajectory of the input $u$ when the proposed model-free ADP technique is applied.

Fig. 7. The trajectory of the input $v$ when the proposed model-free ADP technique is applied.


