Policy-Iteration-Based Adaptive Optimal Control for Uncertain Continuous-Time Linear Systems with Excitation Signals

Jae Young Lee¹, Jin Bae Park¹*, and Yoon Ho Choi²

¹Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea (Tel : +82-2-2123-2773; E-mail: {jyounglee, jbpark}@yonsei.ac.kr)
²Department of Electronic Engineering, Kyonggi University, Suwon, Kyonggi-Do, Korea (Tel : +82-31-249-9801; E-mail: yhchoi@kyonggi.ac.kr)

Abstract: This paper proposes a novel policy-iteration-based adaptive optimal scheme for uncertain continuous-time linear systems with excitation signals. The proposed method can solve the related linear quadratic optimal control problem in online fashion exactly and safely. In order to maintain persistence excitation condition, the controller injects the small excitation signals to the system. For this linear system with excitation signals, the policy iteration (PI) technique is investigated to adaptively find the optimal control law in the presence of both internal uncertainties and known excitation signals. For the proposed PI technique, the stability of the closed-loop system and convergence to the optimal solution are mathematically proven. Numerical simulations are carried out to verify the effectiveness of the proposed method.

Keywords: Policy iteration, adaptive control, LQR, reinforcement learning, excitation condition, PE

1. INTRODUCTION

Policy iterations (PIs) are iterative techniques for solving optimal control problems in forward time, and are extensively applied to finite Markov decision processes in the early stages [1]–[2]. During the last decades, researches on PI and its variants are carried out for discrete-time [3]–[6], and recently, continuous-time [7]–[10] dynamical systems (for recent survey and development, see [11, 12]). Among these schemes, PI schemes proposed in [3, 4, 6, 9, 10, 12] can be considered as reinforcement learning (RL) algorithms which are inspired by biological process for finding optimal behavior from interaction with unknown environments [1]–[2]. In control engineering perspectives, this means that the above PI schemes can be considered as adaptive optimal control schemes, which find the optimal solution in the presence of model uncertainties. In traditional control engineering framework, these kinds of adaptive optimal control strategy are hard to achieve since the conventional adaptive control techniques usually does not guarantee the optimal performance, and optimal control theory can be applied only for perfectly known dynamical systems in offline sense.

However, these PI-based adaptive optimal control schemes require the persistent excitation condition [3, 6, 9, 12], which does not hold anymore if the state almost converges to the desired position and becomes stationary. Without persistent excitation, the controller parameters do not converge to the optimal ones, and in the worst case, numerical instability may be caused due to large computational errors. Moreover, for continuous-time case, the existing methods did not considered the additional excitation signals other than control inputs while the similar probing noise are considered in discrete-time algorithm [6].

In this paper, we propose a PI-based adaptive optimal control scheme for continuous-time uncertain linear systems with excitation signals. The proposed method solves the related linear quadratic optimal control problem in online fashion exactly and safely. In order to maintain persistence excitation condition, the controller injects the small excitation signals to the system, which helps the parameters well-evaluated numerically as well as converge to the optimal solution. In addition, we investigate the novel PI technique, which is the general case of the existing algorithm [9], to adaptively find the optimal control law in the presence of both internal uncertainties and known excitation signals. For the proposed PI technique, the stability of the closed-loop system and convergence to the optimal solution are mathematically proven. Numerical simulations are carried out to verify the effectiveness of the proposed method in comparison to the existing method [9].

2. PRELIMINARIES

In this paper, we consider the infinite horizon LQ OCP for the following system:

\[
\dot{x}(t) = Ax(t) + B[u(t) + w(t)], \quad x(0) = x_0
\]

with the value function

\[
V_u(x(t), t) = \int_t^\infty r(x, u) \, d\tau
\]

for a policy \(u\), where \(x(t) \in \mathbb{R}^n\) and \(u(t) \in \mathbb{R}^m\) are the state and input vectors; \(r(x, u)\) is the quadratic cost defined as \(r(x, u) := x^T Q x + u^T R u\) with \(Q > 0\) and \(R \geq 0\); \(A\) and \(B\) are constant matrices with compatible dimensions; Here, \(w(t)\) denotes the known excitation signal which plays a key role in parameter convergence of the proposed algorithm. Throughout the paper, \(u(t)\) and

* Corresponding author.

This research was supported by the Brain Korea 21 Project in 2010, and by the Ministry of Knowledge Economy (MKE), Project # 2009T100100609, “Development for 22.9kV Power Cable Fault Diagnosis System and Traceability Management System Based on TFDR.”
$u$ will be used interchangeably for the input of the system (1), and so are $x(t)$, $x_i$, and simply $x$ for the state of the system (1) for notational convenience. From the conventional optimal control theory [13], the optimal control law $u^*(t)$ minimizing (2) and its corresponding optimal value function $V^*(x_0)$ for the system (1) with $w(t) \equiv 0$ is given by

$$u^* = -Kx, \quad V^*(x) = x^TPx,$$

where $K := -R^{-1}B^TP$, and $P \geq 0$ is the solution of the algebraic Riccati equation (ARE):

$$A^TP + PA - K^TRK + Q = 0. \quad (4)$$

For the existence and uniqueness of the solution $P \geq 0$, we assume that $(A, B, Q^{1/2})$ is stabilizable and detectable. In the next section, we derive the PI-based adaptive optimal scheme for the system (1) with $w(t) \neq 0$.

In the derivations of the least squares equation of the proposed PI (see (17)), we use the following matrix theory related with Kronecker product and vectorization. Let $A \otimes B$ for the matrices $A$ and $B$ be the Kronecker product of $A$ and $B$. Then, there are three basic properties of Kronecker product, which is needed in this paper:

1) $(A \otimes B)^T = A^T \otimes B^T$
2) $x^TAy = (y \otimes x)^T \text{vec}(A)$
3) vec$(AXB) = (B^T \otimes A) \text{vec}(X)$. 

Here, we denote vec$(X)$ for $m \times n$ matrix $X$ as a vectorization map from a matrix into an $nm$-dimensional real vector. This vec$(X)$ stacks the columns of $X$ on top of one another. Also, we let vec$^*(Y)$ be defined as an operator which maps a symmetric $n \times n$ matrix $Y$ into a vector with dimension $n(n+1)/2$ by stacking the columns corresponding to the diagonal and upper triangular part of $Y$ on top another where upper triangular terms of $Y$ are doubled. For these mapping operators and Kronecker product, we have the fourth property: 4) there always exists a matrix $S$, depending on $n$, such that vec$(Y) = S \text{vec}^*(Y)$ [7].

3. MAIN RESULTS

In this section, we develop a PI-based adaptive optimal control scheme with known exciting signal $w(t)$ for the continuous-time linear system (1). The proposed adaptive optimal control scheme guarantees stability and is able to find the optimal solution $u^*$ and $V^*(x)$ in the presence of the known excitation signal $w(t)$ and uncertainties in $A$.

From now on, we assume that $w(t)$ can be any available signal which is bounded by $w_M$, t.e., $||w(t)|| \leq w_M$. This nonzero $w(t)$ plays a key role in exciting the signals $x_i$ and $u(t)$ for parameter convergence. Based on the existing PI technique [9], the proposed PI, which is able to deal with $w(t)$, is derived as follows:

Policy Evaluation:

$$\Phi_i(t, T) = \int_t^{t+T} x^TP_iBw \, d\tau \quad (5)$$

$$V_i(x_i) + 2\Phi_i(t, T) = \int_t^{t+T} r(x, u_i) \, d\tau + V_i(x_{i+T}) \quad (6)$$

Policy Improvement:

$$u_i(t) = -R^{-1}B^TP_{i-1}x(t) \quad (7)$$

where $V_i(x_i)$ is the $i$-th approximate value function represented by a quadratic form $V_i(x_i) := x_i^TP_i x_i$. Note that if $w(t) \equiv 0$ for all iterations, this PI becomes the existing one [9]. Now, we prove the stability and convergence properties of the proposed PI technique. For notational convenience, define $K_i$ and $A_i$ as follows:

$$K_i := -R^{-1}B^TP_{i-1}, \quad A_i := A - BK_i. \quad (8)$$

With these notations, the closed-loop system at $i$-th iteration is represented by $\dot{x} = A_ix + Bw_i$. Before the proof, we need the following lemma concerned with the stability of the closed-loop system and equivalent matrix equality:

**Lemma 1**: If $A_i$ is stable, the iteration (6)–(7) is equivalent to solving the following Lyapunov equation for $P_i > 0$:

$$(A_i)^TP_i + P_iA_i = -(K_i)^TRK_i - S. \quad (9)$$

**Proof**: Assume that $A_i$ is stable. Then, for $(K_i)^TRK_i + S > 0$, there is $P_i > 0$ such that (9) holds. Considering the Lyapunov function $V_i(x_i) = x_i^TP_i x_i$ and its derivative $\dot{V}_i(x_i) = x_i^T(A_i^TP_i + P_iA_i)x_i + 2u_{i+1}^TR^w$, we have

$$\int_t^{t+T} r(x, u_i) \, d\tau = \int_t^{t+T} x_i^T[S + (K_i)^TRK_i]x_i \, d\tau$$

$$= -\int_t^{t+T} \dot{V}_i(x_i) - 2x_i^TP_iBw \, d\tau$$

$$= V_i(x_i) - V_i(x_{i+T}) + 2\Phi_i(t, T),$$

which completes the proof. \[\square\]

**Theorem 1**: Assume that $(A, B, Q^{1/2})$ is stabilizable and detectable. If the initial controller $u_0$ is stabilizing, and the controller $u_i$ and its value function $V_i(x)$ are updated by the PI scheme (6)–(7), then, the closed-loop system $\dot{x} = A_ix + Bw_i$ is uniformly ultimately bounded (UUB) for all $i \in \mathbb{Z}$, with each ultimate bound:

$$\|x\| \leq \frac{2w_M\|P_{i-1}B\|}{\lambda_m(M_i)\theta_i} \quad (10)$$

for any constant $\theta_i \in (0, 1)$, where $M_i := Q + K_i^TRK_i$. 

Proof: We will prove this by mathematical induction. First, assume that $A_{i-1}$ is stable. Consider the Lyapunov function candidate $V_{i-1}(x_i) = x_i^T P_{i-1} x_i$ for the $i$-th system $\dot{x} = A_i x + B w$. Then, we obtain
\[
\dot{V}_{i-1}(x) = x_i^T (A_{i-1}^T P_{i-1} + P_{i-1} A_{i-1}) x + 2 u_i^T R w
+ x_i^T [P_{i-1} B \Delta K_i + \Delta K_i^T B^T P_{i-1}] x
\]
where $\Delta K_i := K_{i-1} - K_i$. Using Lemma 1 and completing the squares yield
\[
\dot{V}_{i-1}(x) = -x_i^T (S + K_{i-1}^T R K_{i-1}) x + 2 u_i^T R w
- x_i^T [K_i^T R \Delta K_i + \Delta K_i^T R K_i] x
\leq -x_i^T (S + K_{i-1}^T R K_{i-1}) x + 2 u_i^T R w.
\]
Since we assume that $(A, Q^{1/2})$ is observable, $M_i$ is positive definite. So, we have
\[
\dot{V}_{i-1}(x) \leq -x_i^T M_i x + 2 x_i^T P_{i-1} B w
\leq -\lambda_M(M_i) \|x\|^2 + 2 \lambda_M(M_i) \|P_{i-1} B\| \|x\|
= -\lambda_M(M_i) \left( (1 - \theta_i) \|x\|^2 + \theta_i \|x\|^2 - \gamma \|x\| \right).
\]
for some $\theta_i \in (0, 1)$. Here, $\gamma > 0$ denotes $2 \lambda_M(M_i) \|P_{i-1} B\| / \lambda_M(M_i)$. Therefore, $V_{i-1}(x_i) < 0$ holds if $M_i > 0$ and $x$ satisfies $\|x\| \geq \gamma / \theta_i$, which means that by the Lyapunov-like theorem [14], the system $\dot{x} = A_i x + B w$ is uniformly ultimately bounded (UUB) with the ultimate bound (10). Moreover, when $w \equiv 0$, $V_{i-1}(x_i) < 0$ always holds, which means that $A_i$ is stable by Lyapunov theorem [14]. Therefore, if we assume the initial stabilizing controller $u_0$, then, by induction, we prove that the system $\dot{x} = A_i x + B w$ is UUB for all $i \in \mathbb{Z}$, with each ultimate bound (10).

From the above arguments, one can realize that the PI (6)–(7) is actually equivalent to solving the matrix equation (9), provided that the initial policy $u_0$ is stabilizing. From this fact, we obtain the following corollary concerning the stability and convergence of the proposed PI scheme:

**Corollary 1:** Assume that $(A, B, Q^{1/2})$ is stabilizable and detectable. Then, for all $i \in \mathbb{N}$, the PI (6)–(7) conditioned by an initial stabilizing controller $u_0$ yields the stable matrix $A_i$ of the $i$-th closed-loop system, and $K_i$ and $P_i$ eventually converges in second order to the optimal solution $K$ and $P$, respectively.

**Proof:** The stability of $A_i$ is trivially guaranteed by Theorem 1. Now, we show the convergence to the optimal solution. Note that (1) can be expressed as
\[
A_i^T (P_i - P_{i-1}) + (P_i - P_{i-1}) A_i
= -(A_i^T P_{i-1} + P_{i-1} A_i - P_{i-1} B R^{-1} B^T P_{i-1} + Q),
\]
which is equivalent to the Kleinman’s Newton method [15], [9]. Since this method is proven to converge to the optimal solution $P$ in second order [15] and $K_i$ is expressed as $K_i = R^{-1} B P_i$, $P_i$ and $K_i$ converge in second order to $P$ and $K$, respectively, which completes the proof.

Note that $P_i$ contains $N_{min} := n(n+1)/2$ parameters to be estimated by the proposed PI, but there is just one-dimensional equation (6) provided for such calculations. Therefore, we use least squares (LS) method for each step to evaluate all the parameters $P_i$ uniquely. For applying the LA algorithm to the problem, one should modify the formula (6) to a tractable form by using the matrix theory already discussed in Section 2. By using those properties, $x_i^T P_i x + \Phi_i(t, T)$ can be represented as
\[
x_i^T P_i x = ([x \otimes x]^T S) \text{vec}^*(P_i) \quad (11)
\]
\[
= \left[ \int_t^{t+T} (B w \otimes x)^T dr \right] \text{vec}^*(P_i),
\]
Now, using the above expressions and defining $\bar{x}_t := S^T(x_t \otimes x_t)$, (6) can be rewritten as
\[
X^T \text{vec}^*(P_i) = Y, \quad (13)
\]
where
\[
X := \bar{x}_t - \bar{x}_{t+T} + 2 \int_t^{t+T} S^T(B w \otimes x) \, dt, \quad (14)
\]
\[
Y := \int_t^{t+T} x_i^T Q x + u_i^T R u_i \, dt. \quad (15)
\]
Considering the differential equation $\dot{V}(t) = x_i^T Q x + u_i^T R u_i$, $Y$ can be represented as a simple form $Y = V(t+T) - V(t)$. Similarly, we can express $X$ as
\[
X := \bar{x}_t - \bar{x}_{t+T} + 2 \left[ W(t+T) - W(t) \right], \quad (16)
\]
where $W(t)$ is defined by the differential equation $\dot{W} = S^T(B w \otimes x)$.

![Fig. 1 The LS implementation of the proposed PI](image-url)
\[ x^{(j)} \ (1 \leq j \leq N), \text{ respectively, then, vec}^*(P_i) \text{ at } i\text{-th iteration can be obtained from the least squares solution:} \]

\[
\text{vec}^*(P_i) = (X X^T)^{-1}X Y^T, \tag{17}
\]

where \( X := [X^{(1)}, X^{(2)}, \ldots, X^{(N)}]^T \) and \( Y := [Y^{(1)}, Y^{(2)}, \ldots, Y^{(N)}]^T \). In this method, \( N \geq N_{\text{min}} \) is the necessary condition for the excitation condition, namely, the existence of \((X X^T)^{-1}\). One may also use recursive least squares implementation to find \( \text{vec}^*(P_i) \). In this case, persistently exciting condition is needed for parameter convergence at each \( i\)-th iteration.

**Remark 1:** Injecting the small excitation signal \( w(t) \) can prevent the excitation condition from being lost since \( x \) does not converge but excite due to \( w(t) \). If the excitation signal \( w(t) \) is absent, \( x \) eventually converges to zero, implying that the excitation condition (or equivalently, persistently exciting condition) becomes lost. In this case, the LS problem (17) cannot be solved anymore (if not, the solution may contain large errors).

**Remark 2:** By the virtue of the term \( W(t + T) - W(t) \) in (16), \( w(t) \) can also excites each \( X^{(j)} \), which yield a smaller condition number of \( X X^T \). Therefore, the numerical stability concerned about the evaluation \( (X X^T)^{-1} \) can be improved.

### 4. SIMULATION RESULTS

In this section, two numerical simulations are carried out to verify the effectiveness of the proposed PI scheme—one for \( w(t) = 10^{-2} \sin 2 \pi t \), and the other for \( w(t) \equiv 0 \). Note that for the case \( w(t) \equiv 0 \), the PI algorithm (6)–(7) becomes the existing PI proposed in [9]. For both simulations, the system model (1) and the value function (2) are given by

- **System model:**

\[
A = \begin{bmatrix} -6.05 & 1.15 \\ 2.22 & -10 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{18}
\]

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Condition No. with ( w(t) \neq 0 )</th>
<th>Condition No. without ( w(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>103.8</td>
<td>112.5</td>
</tr>
<tr>
<td>2</td>
<td>300.6</td>
<td>9.9479 \times 10^4</td>
</tr>
<tr>
<td>3</td>
<td>14.5</td>
<td>1.7126 \times 10^6</td>
</tr>
<tr>
<td>4</td>
<td>86.8</td>
<td>2.8620 \times 10^8</td>
</tr>
<tr>
<td>5</td>
<td>208.9</td>
<td>4.7660 \times 10^{10}</td>
</tr>
<tr>
<td>6</td>
<td>195.7</td>
<td>7.9158 \times 10^{12}</td>
</tr>
<tr>
<td>7</td>
<td>196.5</td>
<td>4.3852 \times 10^{14}</td>
</tr>
<tr>
<td>8</td>
<td>196.4</td>
<td>2.3802 \times 10^{14}</td>
</tr>
<tr>
<td>9</td>
<td>196.5</td>
<td>3.0404 \times 10^{14}</td>
</tr>
<tr>
<td>10</td>
<td>196.5</td>
<td>4.5374 \times 10^{14}</td>
</tr>
</tbody>
</table>

Table 1 Variations of condition number of \( X X^T \) for both simulations: i) \( w(t) = 10^{-2} \sin 2 \pi t \) and ii) \( w(t) \equiv 0 \).

- **Value function:**

\[
V_u(x(t), t) = \int_t^\infty x_1^2 + x_2^2 + u^2 \, d\tau \tag{19}
\]

with initial condition \( x_0 = [0.3 \ - 0.125]^T \). Here, we denote the system state by \( x := [x_1 \ x_2]^T \). For the above system and value function, the solution \( P \) of the LQ optimal problem can be obtained as follows by directly solving the ARE (4):

\[
P = \begin{bmatrix} 0.0875 & 0.0133 \\ 0.0133 & 0.0514 \end{bmatrix} \tag{20}
\]

Now, we apply the proposed PI scheme (6)–(7) to the system (18) to find the solution \( P \) in (20) in online fashion without any knowledge about \( A \). In this case, at least \( N_{\text{min}} = 3 \) data samples should be collected to solve the LS problem (17). In both simulations, we collect \( N = 5 \ (\geq N_{\text{min}}) \) data per iteration step to evaluate the LS solution.

The simulation results are illustrated in Fig. 2. Note that there are irregular perturbations in \( P \), when \( w(t) \equiv 0 \) after the states go to almost zero (see Figs. 2 (b) and (d)). On the contrary, by virtue of the excitation signal \( w(t) = 10^{-2} \sin 2 \pi t \), no such perturbations arise as shown in Figs. 2 (a) and (c). This is because the state \( x \) is constantly varying by \( w(t) = 10^{-2} \sin 2 \pi t \) and does not become stationary while it is not when \( w(t) \equiv 0 \). Also note that \( w(t) \equiv 0 \) does not deteriorate the regulation performance as can be seen in Fig. 2 (c).

Table 1 shows the condition numbers of \( X X^T \) at each iteration. The condition number implies how much the matrix \( X X^T \) is well-posed. If the number is very large, it implies that the matrix is almost singular and the computation of (17) may introduce large numerical errors [16]. For the case \( w(t) \equiv 0 \), the condition number becomes very large after the state converges, which is illustrated in the second column of Table 1. This introduces the numerical error, and thus, induces the perturbation of \( P_i \) in Figs. 2 (b) and (d). The more state errors occurred from the system and less computational capability, the more perturbations in \( P_i \) may be induced. Moreover, if these perturbations are very large, it may even influence the stability of the closed-loop system. On the contrary, in the case of \( w(t) = 10^{-2} \sin 2 \pi t \), the condition numbers of \( X X^T \) becomes significantly smaller in comparison to those for \( w(t) \equiv 0 \) (see Table 1), which implies the improvement of computational accuracy and solvability, especially when the states almost converge to zero.

In the case of \( w(t) = 10^{-2} \sin 2 \pi t \), the critic parameters of \( P_i \) become stationary and converge to \( P \). The actual final \( P_f \) obtained from the iteration is given by

\[
P_f = \begin{bmatrix} 0.0875 & 0.0133 \\ 0.0133 & 0.0514 \end{bmatrix},
\]

which shows the convergence of the proposed PI as well as the effectiveness of the proposed algorithm.
Fig. 2 Critic parameter variations for (a) $w(t) = 10^{-2} \sin 2\pi t$ and (b) $w(t) \equiv 0$; the state trajectories for (c) $w(t) = 10^{-2} \sin 2\pi t$ and (d) $w(t) \equiv 0$; the trajectory of the control input $u_i + w$ for (e) $w(t) = 10^{-2} \sin 2\pi t$ and (f) $w(t) \equiv 0$.

5. CONCLUSIONS

In this paper, we proposed a novel PI-based adaptive optimal scheme for solving linear quadratic optimal control problems for continuous-time linear systems in the presence of uncertainties in $A$ and the known excitation signal $w(t)$. The excitation signal $w(t)$ played an important role in critic parameter evaluation and its convergence when the proposed PI scheme is applied. The proposed PI scheme was able to deal with this excitation signal, and its stability and convergence to the optimal solution were mathematically proven. Two numerical simulations for $w(t) \equiv 0$ and $w(t) \neq 0$ are carried out to present the effects on $w(t)$ as well as the effectiveness of the proposed method. As a result, if we use this excitation signal $w(t)$, then, the condition number was prevented from begin rapidly growing, which means that the computational accuracy and solvability were improved.
REFERENCES


