2009년도 대한전기학회 하계학술대회 논문집 2009. 7. 14 - 17

안정도 단계가 고려된 LQ 최적 제어에 대한 근사 다이나믹 프로그래밍

이재영*, 박진배*, 최윤호**
연세대학교 전기전자공학과*, 경기대학교 전자공학과**

Approximate Dynamic Programming for Linear Quadratic Optimal Control with Degree of Stability

Jae Young, Lee*, Jin Bae Park*, Yoon Ho Choi**
Dept. of Electrical & Electronic Eng., Yonsei Univ.*, Dept. of Electronic Eng., Kyunggi Univ.**

Abstract - 본 논문에서는 안정도 단계(degree of stability)가 고려된 LQ 최적 제어에 대한 근사 다이나믹 프로그래밍 기법을 제안한다. 제안된 근사 다이나믹 프로그래밍 기법은 시스템 행렬(system matrix)를 모르는 경우에도 구현할 수 있으며, 특정 조건하에서 수렴성을 가짐을 수학적으로 증명하였다. 또한 제안된 알고리즘을 토대로 하는 최소 자승법 기반 실시간 구현 방법에 대해 소개하였으며, 컴퓨터 모의 실험을 통해 제안된 근사 다이나믹 프로그래밍의 성능을 입증하였다.

1. 서 론

근사 다이나믹 프로그래밍 (approximate dynamic programming: ADP)은 강화 학습 기반의 반복계산을 통해 시스템에 대한 최적 입력을 구하는 방법을 일컫는다[1]. 이산시간 시스템과 연속시간 시스템에 대한 이러한 ADP에 대한 연구가 지속적으로 진행되고 있다. 특히 이산시간 기반 ADP의 종류 중 하나인 ADHDP (Q-learning) 기법은 모델을 모르는 상태에서도 최적입력을 구할 수 있다는 점에서 다른 ADP 기법에 비해 더욱 주목받는 방법이다[2]. 하지만, 많은 응용 연구들이 수렴성에 대한 고려를 하지 않은 채로 ADHDP를 사용해 왔고, 이는 시스템의 불안정성을 초래하는 원인이 되었다. 이를 해결하기 위해 Bradke는 이산시간 LQ 최적 제어에 대한 ADHDP 기법의 수렴성을 증명하였다[3]. 한편, Baird에 의해 연속시간에서는 Q-learning을 사용할 수 없다는 사실이 지적되었고[4], 이에 따라 연속시간에서의 Q-learning을 대체할 수 있는 ADP 알고리즘들이 개발되었다[4-5]. 하지만, [4-5]에서 개발된 알고리즘 역시 수렴성을 보장하지 못한다는 문제를 가지고 있다. Murray, Vrabie 등은 시스템 행렬(system matrix)을 모르는 선형 시스템에 대해 적용 가능한 ADP 알고리즘을 제안하였다 [6]. 제안된 알고리즘은 비록 입력 결합 행렬을 알아야 한다는 제약이 있지만, 수렴성이 어느 정도 보장된 기법들로, 기존 알고리즘이 가지고 있는 단점을 크게 보완하였다.

한편 강화학습 기법에서는 현재와 미래의 보상값에 가중치를 두는 방법을 도입하여 학습 방법을 다양화시켰다 [7]. 이렇게 현재와 미래의 비용에 가중치를 주는 기법은 제어공학 영역에서도 개발되었는데, 이러한 가중치 방법을 사용하면, 최적 제어 시스템에 안정도 여유를 줄 수 있다는 장점이 있다 [8]. 하지만, 이러한 가중치 방법의 장점에도 불구하고, 그에 대한 ADP 방법에 대한 연구는 전무한 실정이다. 본 논문에서는 이러한 가중치, 즉 안정도 단계(degree of stability)가 고려된 2차 비용함수와, 불확실한 시스템 행렬을 갖는 연속시간 선형시스템에 대한 ADP 기법을 제안한다. 제안된 방법을 통해 안정도 단계가 고려된 LQ 최적해를 구할 수 있으며, 제안된 기법이 일정 조건하에서 수렴성을 가짐을 본 논문에서 증명하였다. 또한, 제안된 방법을 토대로 하는 최소자승법 기반 ADP 알고리즘 구현 방법을 소개하고, 컴퓨터 모의실험을 통해 제안된 방법의 성능을 입증하고자 한다.

2. 안정도 단계가 고려된 ADP 알고리즘

제어 대상이 다음과 같은 선형 시스템으로 표현된다고 가정하자.

$$\dot{x} = Ax + Bu,\ x(0) = x_0 \tag{1}$$

여기서 $x \in R^n$와 $u \in R^m$은 각각 시스템의 상태변수와 제어입력이고, A와 B는 상수계수를 갖는 행렬이다. 위 시스템에 대해 다음과 같은 성능 함수를 고려하자.

$$V(x(t)) = \int_t^\infty e^{2\alpha(\tau - t)}(x^T Q x + u^T R u)d\tau \tag{2}$$

여기서 α는 안정도 단계를 나타내는 상수이고, Q 와 R 은 각각 양확정(positive definite), 준양확정(positive semi-definite) 행렬이다. 본 논문에서 제안하는 ADP 알고리즘의 목적은 행렬 A에 대한 정보가 주어지지 않았을 경우 상기의 비용 함수 V를 최소화 시키는 제어 입력 $u^* = K^* x$를 구하 것이다. 그러한 행렬 K^*의 존재성 및 유일성을 위해 시스템 (1)에 대해 다음과 같이 가정한다.

가정 1: $(A, B, Q^{1/2})$는 안정화 가능(stabilizable)하고 측정 가능(detectable) 하다.

위 가정하에 최적행렬 K^*는 행렬 A와 B를 알고 있는 경우 다음과 같이 구할 수 있다.

$$K^* = -R^{-1}B^T P \tag{3}$$

여기서 준양확정 행렬 P는 아래와 같은 리카티 대수 방정식(algebraic Riccati equation: ARE)을 만족시킨다.

$$Ric(P) := 2\alpha P + A^T P + PA - PBR^{-1}B^T P + Q = 0 \tag{4}$$

상기 무한 계획(infinite horizon) 성능 함수 V는 다음과 같이 근사화 시킬 수 있다.

$$V(x(t)) = \int_t^{t+T} e^{2\alpha(\tau - t)}(x^T Q x + u^T R u)d\tau + e^{2\alpha T} W(x(t+T)) \tag{5}$$

여기서 $W(x(t))$는 $V(x(t))$의 근사함수이다. 식 (5)를 기반으로 반복 알고리즘을 도출하면 다음과 같다.

$$V_{i+1}(x(t)) = \int_t^{t+T} e^{2\alpha(\tau - t)}(x^T Q x + u^T R u)d\tau + e^{2\alpha T} V_i(x(t+T)) \tag{6}$$

여기서 i는 반복횟수를 나타내고, $V_i(x)$는 i번째 반복점에서 구한 성능함수의 근사함수이다. 위 식에서 $V_i(x)$는 $V_i(x) = x^T P_i x$ $(P_i \geq 0)$와 같은 이차형식으로 나타낼 수 있으므로, 식 (3)과 식 (6)을 기반으로 다음과 같은 ADP 반복 알고리즘을 얻는다.

성능함수 업데이트:

$$x^T(t)P_{i+1}x(t) = \int_t^{t+T} e^{2\alpha(\tau - t)}(x^T Q x + u^T R u)d\tau + e^{2\alpha T}x^T(t+T)P_i x(t+T) \tag{7}$$

제어기 업데이트: $u_i(x) = K_i x$ (8)

여기서 $K_i := -R^{-1}B^T P_i$로 정의된다. 상기의 ADP 알고리즘에는 행렬 A가 포함되지 않았으므로, 행렬 A를 모르더라도 위 반복법에 의해 최적 입력 u^*를 얻을 수 있다.

Remark 1: $\alpha = 0$으로 놓으면, 위 알고리즘 (7)-(8)은 Vrabie의 ADP 알고리즘[6]과 같아진다. 따라서 Vrabie의 알고리즘은 본 논문에서 제안된 알고리즘의 특수한 형태라 할 수 있다.

정리 1: 만일 $P_0 \geq 0$ 이면, 모든 i 에 대해 $P_i \geq 0$ 가 성립한다.
증명: 지면 제한상 본 정리의 증명은 생략한다.

정리 2: 위 선형 시스템 (1)이 가정 1을 만족하고, 행렬 $A + BK_0 + \alpha I$가 Hurwitz라고 가정하자. 만일 알고리즘 (7)-(8)이 매 i번째 단계에서 부등식

$$\frac{2\overline{\lambda_i}}{T\|BR^{-1}B'\|\|Ric(P_i)\|} \geq 1 \tag{9}$$

을 만족시킨다면, 모든 i에 대해서 행렬 $A + BK_i + \alpha I$가 Hurwitz이고, 다음이 성립한다.

1) $\|Ric(P_{i+1})\| < \|Ric(P_i)\|$,
2) $\lim_{i \to \infty} \|Ric(P_i)\| = 0$

여기서 $-\overline{\lambda_i}$는 행렬 $A + BK_i + \alpha I$의 최대 고유값이다.
증명: 지면 제한상 본 정리의 증명은 생략한다.

정리 2는 제안된 ADP 방법의 수렴성에 관한 정리이다. 상기 조건 (9)를 만족시키기 위해서는 행렬 R의 노름 $\|R\|$과 오차 $\|Ric(P_i)\|$, 주기 T가 충분히 작아야만 한다.

3. 최소자승법 기반 알고리즘 구현 및 시뮬레이션

본 논문의 2장에서 제안된 ADP 알고리즘은 매 i번째 반복시행에서 P_i의 모든 항목을 변경시킨다. 즉, $N_{\min} := (n+m)(n+m+1)/2$ 개의 변수가 한번에 갱신(update) 된다. 하지만, 그와 관련된 식은 단 하나의 식 (7)만이 주어져 있기 때문에 알고리즘 실행에 문제가 발생한다. 본 절에서는 이를 해결하기 위한 최소자승법 기반 ADP 알고리즘을 소개하고, 이를 이용한 시뮬레이션 및 그 결과에 대해 토의한다.

3.1 최소자승법 기반 ADP 알고리즘

위와 같은 문제를 해결하기 위해서는 T를 주기로 매 순간의 상태정보를 획득해야 한다. i번째 반복시행에서 획득한 $N \geq N_{\min}$개의 상태정보를 각각 $x_1^{(i)}$, $x_2^{(i)}$, ..., $x_N^{(i)}$라 하자. 그러면, 각각의 $x_k^{(i)}$, (k = 1,2,..., N)에 대한 i번째 성능함수 $V_i(x)$는 다음과 같이 표현된다.

$$V_i(x_k^{(i)}) = (x_k^{(i)})^T P_i x_k^{(i)} = (\bar{x}_k^{(i)})^T \varrho(P_i) \tag{10}$$

여기서 $\bar{x}_k^{(i)}$는 $x_1^{(i)}$의 크로네커 곱(Kronecker product)이고, $\varrho(\cdot)$는 $n \times n$ 대칭행렬(symmetric matrix)을 $N_{\min}$차원의 열벡터로 사상시키는 연산으로, 다음과 같이 정의된다.

$$\begin{aligned}\varrho(P) = [&[p_{11}, p_{21}+p_{12}, p_{31}+p_{13}, \cdots, p_{(n+m)1}+p_{1(n+m)}],\\ &[p_{22}, p_{32}+p_{23}, \cdots, p_{(n+m)2}+p_{2(n+m)}],\\ &\vdots\\ &[p_{(n+m)(n+m)}]]^T\end{aligned} \tag{11}$$

여기서 p_{jk}는 행렬 P의 (j,k)번째의 원소이다. 이제, 위 표현들을 바탕으로 식 (7)을 다음과 같이 나타낼 수 있다.

$$(\bar{x}_k^i)^T \varrho(P_{i+1}) = V(t+T) - V(t) + e^{2\alpha T}(\bar{x}_{k+1}^i)^T \varrho(P_i) \tag{12}$$

여기서 $V(t)$는 다음과 같은 시스템으로 정의된다.

$$\dot{V} + 2\alpha V = x^T Q x + u^T R u, \quad V(0) = 0. \tag{13}$$

위 시스템 (13)에서 α는 시스템의 대역폭을 결정해 주는 인자(factor)로 작용한다. 이는 실제 시스템에서의 고주파 잡음을 제거하는 역할을 함으로써 잡음에 의한 시스템의 성능 저하를 방지한다. 이제 $d(\bar{x}_k^{(i)}, P_i) := V(t+T) - V(t) + e^{2\alpha T}(\bar{x}_{k+1}^{(i)})^T \varrho(P_i)$로 정의하면, 상기 (12)의 최소 자승해는 다음과 같이 표현된다.

$$\varrho(P_{i+1}) = (X^{(i)} X^{(i)T})^{-1} X^{(i)} Y^{(i)} \tag{13}$$

여기서 $X^{(i)}$와 $Y^{(i)}$는 각각

$X^{(i)} := [x_1^{(i)}, x_2^{(i)}, ..., x_N^{(i)}]$,

$Y^{(i)} := [d(\bar{x}_1^{(i)}, P_i), d(\bar{x}_2^{(i)}, P_i), ..., d(\bar{x}_N^{(i)}, P_i)]^T$

로 정의된다.

3.2 모의실험 결과

본 절에서는 다음과 같은 선형시스템에 대한 모의실험을 통해 본 논문에서 제안된 ADP의 성능을 검증한다. 모의실험에 사용될 제어대상은 다음과 같다.

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}-1 & 0 & 2\\ 1 & -2 & 0\\ 0 & 0 & 3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}0\\ 0\\ 1\end{bmatrix} u \tag{14}$$

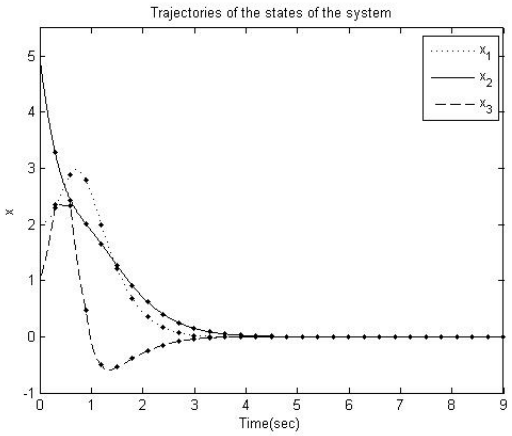
ADP를 통해 최소화시킬 성능함수는 다음과 같다.

$$V(x(0)) = \int_0^\infty e^{2\tau}(x_1^2 + x_2^2 + 0.5x_3^2 + 10^{-2}u^2) d\tau \tag{15}$$

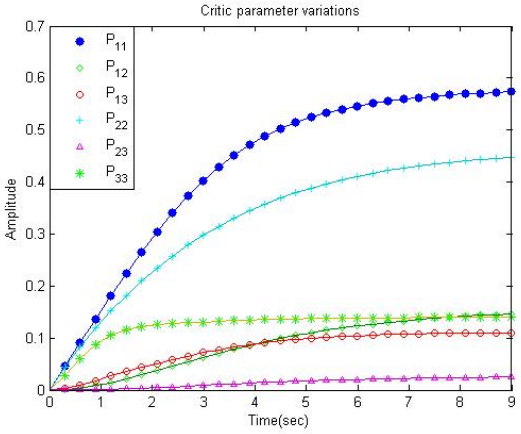
본 모의실험에서는 초기값을 $(x_1, x_2, x_3)^T = (5,2,1)^T$로 설정하였고, 식 (7)의 주기 T를 $T = 50\,[ms]$로 설정하였다. 차수가 3차이므로, 최소 $6T = 300\,[ms]$ 마다 ADP 기반 업데이트가 수행된다. 실험결과는 그림 1, 2와 같다. 위 시스템에 대한 행렬 P의 참값과 본 실험을 통해 구한 행렬 P의 추정치 $\hat{P}$는 다음과 같다.

$$P = \begin{bmatrix}0.5920 & 0.1557 & 0.1145\\ 0.1557 & 0.4638 & 0.0269\\ 0.1145 & 0.0269 & 0.1457\end{bmatrix},\ \hat{P} = \begin{bmatrix}0.5723 & 0.1452 & 0.1102\\ 0.1452 & 0.4454 & 0.0246\\ 0.1102 & 0.0246 & 0.1398\end{bmatrix}$$

상기 결과로부터 매우 정확하게 추정된 것을 알 수 있다.



〈그림 1〉 상태변수의 궤적



〈그림 2〉 행렬 P_i의 궤적

4. 결 론

본 논문에서는 안정도 단계, 즉 가중치가 고려된 LQ 최적 제어에 대한 ADP 기법을 제안하였다. 제안된 방법은 시스템 행렬을 모르는 경우에도 적용 가능하며, 일정 조건하에서 수렴성을 가짐이 증명되었다. 또한, 이에 대한 최소자승법 기반 ADP 알고리즘 구현 방법이 소개되었고, 모의실험을 통해 제안된 방법의 성능을 입증하였다.

감사의 글

이 논문은 2009년도 교육인적자원부 BK21사업의 일환인 연세대학교 전기전자공학부 TMS사업단과 산업자원부 전력기반조성사업 센터의 고급인력양성사업을 통한 연세 대학교 계통적용 신전력기기 연구센터의 지원을 받아 연구되었습니다.

[참 고 문 헌]

[1] J. Si, A. G. Barto, W. B. Powell, and D. Wunsch, *Handbook of Learning and Approximate Dynamic Programming*, Wiley-IEEE Press, 2004.

[2] D. V. Prokhorov and D. C. Wunsch, II, "Adaptive critic designs," *IEEE Trans. Neural Networks*, vol. 8, no. 5, pp. 997-1007, 1997.

[3] S. J. Bradtke and B. E. Ydstie, "Adaptive linear quadratic control using policy iteration," in *Proc. of American Control Conference*, Baltimore, Maryland, 1994, pp. 3475-3479.

[4] L. C. Baird, III, "Reinforcement learning in continuous time: Advantage updating," in Int. Conf. Neural Networks, Orlando, FL, 1994, vol. 4, pp. 2448-2453.

[5] K. Doya, "Reinforcement learning in continuous-time and space," *Neural Computation*, vol. 12, no. 1, pp. 219-245, Jan. 2000.

[6] D. Vrabie, M. Abu-Khalaf, F. L. Lewis, and Y. Wang, "Continuous-time ADP for linear systems with partially unknown dynamics," in *Proc. of IEEE Int. Symp. ADPRL*, pp. 247-253.

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

[8] D. D. Ruscio, "On the location of LQ-optimal closed-loop poles," in *Proc. of Conf. on Decision and Control*, Brighton, 1991, pp. 1554-1556.