온라인 강화학습 기반 서보 시스템의 최적 각도 추종 제어
Online Reinforcement Learning for Optimal Tracking in Servo Positioning Systems

○이 효 찬[1], 최 경 환[1*]
[1)] 한국과학기술원 조천식모빌리티대학원 (TEL:042-350-1764; E-mail: hyochanlee@ kaist.ac.kr, kh.choi@kaist.ac.kr)

Abstract This paper proposes an online reinforcement learning-based optimal tracking control method for output feedback servo systems under unknown disturbances. The system is reformulated in a control-affine form with uncertainties represented as a lumped nonlinear function. An online identifying filter estimates unknown dynamics in real time, while an actor-critic neural network approximates the value function and optimal control policy. The method yields an approximate Hamilton–Jacobi–Bellman (HJB) solution, with convergence ensured via adaptive learning laws. Simulation results demonstrate robust tracking performance under time-varying disturbances.

Keywords Optimal control, Learning based control, Reinforcement learning, Servo positioning

1. 서론

공정 자동화, 의료 수술용 로봇 등의 로보틱스 응용 시스템은 구동부의 정밀한 서보 시스템 설계를 요구한다. 그러나 실제 시스템은 파라미터 불확실성과 외부의 불확실한 외란 등으로 인해 정확한 궤적 추종이 어려워지는 문제가 발생하며, 이는 전체 시스템의 제어 성능을 크게 저하한다 [1].

이런 문제 해결을 위해 기존에는 피드백 선형화, 백스테핑, 슬라이딩 모드 제어, 외란 관측기 기반 제어 등 시스템의 불확실성에 대응하는 강인성을 중점으로 다양한 비선형 제어 기법이 개발되었으나, 시스템의 성능과 효율을 모두 고려한 최적성에 대한 엄밀한 분석은 부족했다.

본 논문에서는 출력 피드백 기반의 서보 시스템을 대상으로, 실시간으로 현재 상태에 대해 최적 제어 정책을 학습하는 온라인 강화학습 기반 추종 제어 기법을 제안한다. 제안된 방법은 서보 시스템 모델을 기반으로 개루프 시스템을 설계하고, 온라인 필터 설계를 통한 비선형 함수의 추정 및 액터-크리틱 신경망 구조를 통한 최적 가치함수와 최적 제어 솔루션을 근사하여 제어에 활용한다.

2. 강화학습 기반 최적 추종 제어 기법

2.1 서보 시스템 모델

서보 시스템 모델의 전압 v_s (V)에 따른 각도 θ(rad), 속도 ω(rad/s), 전류 $i_s(A)$ 거동 방정식은 다음과 같다.

$$\frac{d\theta}{dt} = \omega, \tag{1}$$

$$J\frac{d\omega}{dt} = -B\omega + T_e(i_s) - T_L, \tag{2}$$

$$L\frac{di_s}{dt} = -Ri_s - b_F(\omega) + v_s, \tag{3}$$

이때 T_L(Nm)은 외부 부하 토크, $T_e(i_s) = k_T i_s$(Nm)는 토크 상수 k_T에 대한 출력 토크, $b_F(\omega) = k_e\omega$는 역기전력 상수 k_e에 대한 역기전력이다. 나머지 파라미터 계수들은 다음과 같이 정의된다; J: 관성계수(kgm^2), B: 회전자 댐핑 계수($Nm/rad/s$), L-고정자 인덕턴스(H), R-고정자 저항(Ω).

출력 토크와 역기전력의 관계를 활용하여 아래의 개 루프 시스템에 대한 설계가 가능하다.

$$\dot{x} = g_0 u + f, \tag{4}$$

이때, $x = \theta, u = v_s, g_0 = \frac{1}{k_{e,0}}$는 각각 출력 피드백, 전압 입력, 역기전력 상수의 공칭 값이고 $f = -\frac{JR}{k_T k_e}\frac{d\omega}{dt} - \frac{BR}{k_T k_e}\omega - \frac{L}{k_e}\frac{di_s}{dt} - \frac{1}{\Delta k_e}v_s - \frac{R}{k_T k_e}T_L$ 은 불확실한 비선형 함수이다.

2.2 강화학습 기반 최적 추정 제어 기법

제어 추종 오차 $e(t) = x(t) - r(t)$에 대한 보상함수를 $r(e,u) = e^2 + g_0 u^2$와 같이 정의할 때 무한시간 적분에 대한 가치함수는 아래와 같다.

$$V(e) = \int_t^\infty r\big(e(s), u(s)\big)ds, \tag{5}$$

이에 대한 최적 가치함수 $V^*(e)$는 (6)을 만족하며,

$$V^*(e) = \min\left(\int_t^\infty r\big(e(s), u(s)\big)ds\right) = \int_t^\infty r\big(e(s), u^*(s)\big)ds, \tag{6}$$

이를 기반으로 (7)의 HJB 방정식과 최적제어 입력 (8)의 유도가 가능하다.

※이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-00554087).

$$H\left(e,u^*,\frac{dV^*}{de}\right)=r(e,u^*)+\frac{dV^*}{de}(g_0u^*+f-\dot{r}), \quad (7)$$

$$u^*=-\frac{1}{g_0}\left(\gamma_e e+f+\frac{1}{2}V_0\right), \quad (8)$$

이때, $V_0(e,x)=\frac{dV^*}{de}-2\gamma_e-2f$이다.

최적 제어 입력 (8)의 비선형 f를 근사하기 위한 온라인 필터는 아래와 같다.

$$\dot{\hat{x}}=g_0u^*+\hat{f}+\gamma_{f,1}\tilde{x}, \quad (9)$$

$$\hat{f}=\zeta+\gamma_{f,2}\tilde{x}, \quad (10)$$

$$\dot{\zeta}=-\gamma_{f,2}\zeta-\gamma_{f,2}^2\tilde{x}+\gamma_{f,2}\left(\hat{f}+\gamma_{f,1}\tilde{x}\right), \quad (11)$$

이때 $\tilde{x}=x-\hat{x}$은 출력 피드백 상태의 추정 오차 $\gamma_{f,1}>0,\gamma_{f,2}>1$는 필터의 학습률이다.

최적 제어 입력 (8)의 V_0를 근사하기 위한 액터-크리틱 신경망 구조는 아래와 같다.

$$\frac{dV^*}{de}=2\gamma_e e+\hat{f}+W_c^T\sigma_V(e,x), \quad (12)$$

$$\dot{\hat{W}}_c=\gamma_c\sigma_V(e,x)\sigma_V^T(e,x)W_c, \quad (13)$$

$$\hat{u}^*=-\frac{1}{g_0}\left(\gamma_e e+\hat{f}+\frac{1}{2}\hat{W}_a^T\sigma_V(e,x)\right), \quad (14)$$

$$\dot{\hat{W}}_a=-\sigma_V(e,x)\sigma_V^T(e,x)\left(\gamma_a\left(\hat{W}_a-\hat{W}_c\right)+\gamma_c\hat{W}_c\right), \quad (15)$$

이때 W_c, $W_a,\sigma_V(e,x)$는 각각 크리틱, 액터의 가중치와 활성 함수이며, γ_c,γ_a는 각각 크리틱과 액터의 학습률이다.

3. 모의 실험 검증

3.1 모의 실험 환경

시스템 모델 (1)~(3)을 기반으로 MATLAB/Simulink 환경에서 진행되었으며 설정된 모델의 파라미터와 튜닝된 학습률은 다음과 같다: $J=4.3\times10^{-5},B=0.1J,k_T=k_e=0.068,R=0.78,\ L=1.3\times10^{-4},\gamma_e=60,\gamma_{f,1}=100,\gamma_{f,2}=250,\gamma_c=42,\gamma_a=23.$

액터, 크리틱 신경망의 가중치 초기값은 $\hat{W}_a(0)=[6,\cdots,6]^T\in R^{16\times1},\hat{W}_c(0)=[4,\cdots,4]^T\in R^{16\times1}$와 같이 설정되었고, 액터와 크리틱 신경망의 i번째 활성함수($i=1,\cdots,16$)는 아래와 같다.

$$\sigma_{v,i}(e,x)=\exp\left(-\frac{\left([e,x]^T-\frac{1}{2}[i,i]^T\right)^T\left([e,x]^T-\frac{1}{2}[i,i]^T\right)}{3^2}\right)$$

3.2 모의 실험 결과

추종 성능 검증을 위해 $r=8\sin(2\pi\cdot0.3t)+3\cos(2\pi\cdot0.3t)\cos(2\pi\cdot0.3t)\left(1-e^{-0.2t}\right)$의 제어 지령과 $T_L=0.4\theta+\sin^2(2\pi\cdot0.3t)$의 외부 토크가 인가되었으며, 학습 기반 제어기 2 종과 외란 추정 기반 제어기 1 종 간의 비교 검증을 진행하였다. 그림 1은 각도 추종 제어 거동을 나타내며, 그림 2 는 각도 추종 오차를 나타낸다 [2]-[3].

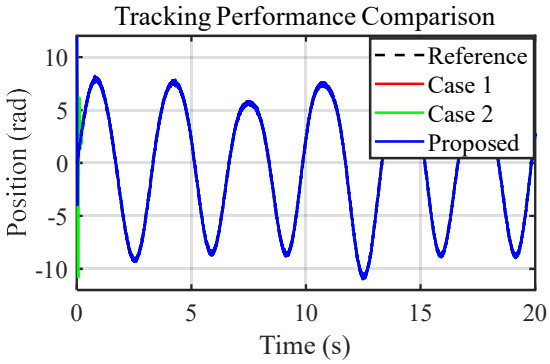

그림 1. 각도 추종 제어 거동

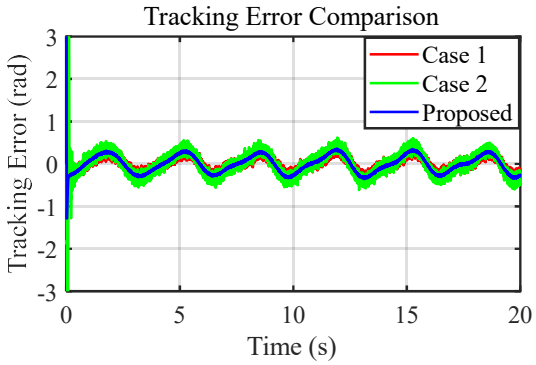

그림 2. 각도 제어 추종 오차

4. 결론

본 논문에서는 출력 피드백 서보 시스템의 각도 추종 제어를 위한 온라인 강화학습 기반 최적 제어 기법을 제안하였다. 제안한 방법은 별도의 모델 식별이나 오프라인 학습 없이 시스템의 불확실한 비선형 동역학을 온라인 필터로 추정하고, 액터-크리틱 신경망 구조를 통해 최적 가치함수와 최적 제어 입력을 근사한다. 모의 실험 검증 결과 제안한 방법은 불확실한 시변 외란 환경에서 실시간으로 최적의 추종 제어 입력을 근사하며 안정적인 제어 성능을 가지는 것을 확인하였다. 추후 실험 환경에서의 검증 계획을 가지고 있다.

참고문헌

[1] Y. Zuo, J. Mei, C. Jiang, X. Yuan, S. Xie, and C. H. T. Lee, "Linear Active Disturbance Rejection Controllers for PMSM Speed Regulation System Considering the Speed Filter," *IEEE Trans on Power Electronics*, vol. 36, no. 12, pp. 14579–14592, Feb. 2021.

[2] G. Wen, C. L. P. Chen, and S. S. Ge, "Simplified Optimized Backstepping Control for a Class of Nonlinear Strict-Feedback Systems With Unknown Dynamic Functions," *IEEE Trans on Cybernetics*, vol. 51, no. 9, pp. 4567–4580, Sept. 2021

[3] I. Lee, S. You, W. Kim, and J. Moon, "Extended state observer–critic architecture-based output-feedback optimized backstepping control for permanent magnet synchronous motors," *Expert Systems with Applications*, vol. 270, p. 126542, Apr. 2025.