온라인 강화학습을 위한 벨만 최적 방정식의 제약 최적화 문제
Constrained Optimization Formulation of Bellman Optimality Equation for Online Reinforcement Learning

○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 a constrained optimization-based reinforcement learning (RL) method to enhance learning stability for discrete-time nonlinear systems. To address the oscillatory convergence and hyperparameter sensitivity of conventional Adaptive Dynamic Programming (ADP), the Bellman Optimality Equation (BOE) is reformulated as a constrained optimization problem. By treating the control policy and critic weights as simultaneous decision variables and enforcing the BOE as an equality constraint, the framework ensures consistent optimality throughout the learning process. Online update laws are derived from the Karush-Kuhn-Tucker (KKT) conditions within a Lagrangian framework to seek the optimal solution in real time. The proposed approach demonstrates improved convergence and robustness over standard RL-based control methods.

Keywords Reinforcement Learning, Constrained Optimization, Bellman Optimality Equation, Optimal Control

1. 서론

강화학습과 근사 동적 계획법은 복잡한 비선형 시스템의 최적 제어 문제를 해결하기 위한 방법론으로 널리 활용되고 있다. 이러한 접근은 벨만 최적 방정식과 같은 최적성 조건을 반복적으로 근사함으로써, 정확한 분석적 해를 구하거나 별도의 사전 계산 과정 없이도 최적의 제어 정책을 자율적으로 학습할 수 있게 한다. 이러한 특징은 시스템 동역학이 불완전하게 알려져 있거나 불확실성이 존재하는 환경에서 운용되는 시스템의 최적 성능을 찾는 데 있어 매우 효과적이다[1].

그러나 정책 반복이나 가치 반복을 기반으로 하는 기존의 프레임워크는 증분형 업데이트 방식이 비선형 함수 근사 기법과 결합될 경우, 수렴을 보장하는 수축 특성이 상실되거나 과도한 분산이 발생하여 학습이 발산할 위험이 나타나기 쉽다. 이러한 불안정성은 수렴 과정에서의 심한 진동이나 완전한 발산으로 이어지며, 결과적으로 제어 시스템이 학습률의 변화나 데이터 분포의 전이에 극도로 민감해지는 원인이 된다[2].

본 논문에서는 이러한 학습 불안정성을 해결하기 위해 벨만 최적 방정식을 제약 최적화 문제로 정형화하여 가중치를 학습하는 새로운 프레임워크를 제안한다. 제안된 방법은 가치 함수 파라미터와 제어 정책을 결합 결정 변수로 설정하고 온라인 학습 알고리즘을 통해 실시간으로 KKT 조건의 해를 탐색한다. 이를 통해 학습 과정 전반에 걸쳐 벨만 오차를 안정적으로 줄여 나감으로써, 기존의 비제약 기반 방식보다 안정하게 최적해를 근사 할 수 있다.

2. 서론

2.1 문제 정의

본 논문에서는 다음과 같은 일반화된 이산시간 비선형 시스템을 고려한다.

$$x_{k+1} = f(x_k) + g(x_k)u_k, \tag{1}$$

비선형 시스템 (1)에 대한 최적 제어 문제는 아래의 무한 시간에 대한 목적 함수 $J(x_k)$를 최소화하는 최적 피드백 정책 u_k^*을 찾는 문제로 정의할 수 있다.

$$J(x_k) = \sum_{i=k}^{\infty} \gamma^{i-k} r(x_i, u_i), \tag{2}$$

이때 γ 는 무한 시간 목적 함수의 할인 인자이며, $r(x_i, u_i)$는 단계 비용 함수이다. 최적 정책 u_k^*에 대한 최적 목적함수 $J^*(x_k)$ 는 벨만 최적 원리에 따라 아래의 벨반 최적 방정식을 만족한다.

$$J^*(x_k) = \min_{u_k}\{r(x_k, u_k) + \gamma J^*(x_{k+1})\}, \tag{3}$$

$$u^*(x_k) = \operatorname*{argmin}_{u_k}\{r(x_k, u_k) + \gamma J^*(x_{k+1})\}, \tag{4}$$

2.1 제약 최적화 문제

최적 가치함수 $J^*(x_k)$를 근사하기 위해 신경망 근사 $J^*(x_k) = W^{*T}\phi(x_k) + \epsilon(x_k)$ 를 통해 이상적인 신경망 가중치 W^*를 찾는 문제로 최적 제어 문제를 변환하여 아래와 같은 제약 최적화 문제를 정의한다.

$$\min_{u_k, W_k} \; r(x_k, u_k) + \gamma \hat{f}(x_{k+1}, W_k)$$
$$s.t. \quad \delta_k(u_k, W_k) = 0, \qquad (5)$$

이때 $\delta_k(u_k, W_k)$는 벨만 최적 방정식의 잔차 오류로 아래와 같이 정의된다.

$$\delta_k(x_k, u_k) = r(x_k, u_k) + \gamma W_k^T \phi(x_{k+1}) - W_k^T \phi(x_k), \quad (6)$$

2.2 제약 최적화 기반 온라인 학습 알고리즘

정의된 제약 최적화 문제(5)를 기반으로 라그랑주 함수는 아래와 같이 정의된다.

$$L = r(x_k, u_k) + \gamma \hat{f}(x_{k+1}, W_k) + \lambda_\delta \delta_k(u_k, W_k), \quad (7)$$

라그랑주 함수 (7)을 기반으로 제약 최적화 문제 (5)를 풀기 위한 온라인 강화학습 알고리즘은 아래와 같이 정의된다.

$$W_{k+1} = W_k - \alpha_W(\nabla_W L + \mu \zeta_k \delta_k), \qquad (8)$$
$$\nabla_W L = \phi(x_{k+1}) + \lambda_{\delta,k} \zeta_k, \qquad (9)$$
$$\lambda_{\delta,k+1} = \lambda_{\delta,k} + \mu \delta_k, \qquad (10)$$
$$u_k = -\frac{\gamma}{2} R^{-1} W_k^T \nabla \phi(x_{k+1}) g(x_k), \qquad (11)$$

이때 ζ_k는 $\zeta_k = \gamma\phi(x_{k+1}) - \phi(x_k)$로 정의된 신경망 기저함수의 회귀 벡터이며, α_W, μ는 각각 신경망의 학습률이다.

3. 실험 검증

학습 성능 검증을 모바일 로봇의 위치 지령에 대한 제어 실험을 진행하였으며, 비교를 위해 TD(0) 학습 기반 알고리즘[3]과 온라인 ADP 기반 알고리즘[4]와의 비교 검증을 진행하였다 그림 1 은 실험 검증에 활용한 모바일 로봇을 나타낸다.

그림 1. 실험 검증 모바일 로봇

모바일 로봇의 위치 지령을 $[x_r, y_r]^T = [2, -1]^T$로 설정하여 실험을 진행하였으며 위치 오차 $d = \sqrt{(y_r - y)^2 + (x_r - x)^2}$ 와 각도 오차 $\theta_e = \tan^{-1}\left(\frac{y_r - y}{x_r - x}\right) - \theta$를 기반으로 아래와 같은 기저 함수를 활용하였다.

$$\phi(x_k) = [d^2, \theta_e^2, d\theta_e, 1 - \cos(\theta_e), d\sin(\theta_e)]^T, \quad (11)$$

그림 2 는 3 번의 에피소드를 반복하였을 때의 모바일 로봇의 위치 거동과 가중치의 거동을 나타낸다.

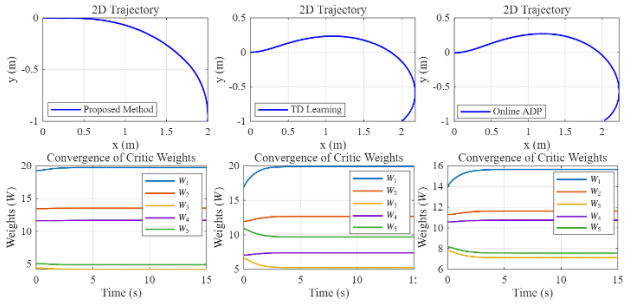

그림 2. 모바일 로봇의 에피소드 3 번 반복에 대한 각 알고리즘의 위치 거동과 가중치 수렴 결과

그림 2 의 위치 거동 실험 결과는 제안 방법의 온라인 학습 결과가 주어진 위치 지령에 대해 안정적인 궤적을 가지며 이동하는 결과를 보여주며, 가중치 수렴 결과는 제안 방법이 안정적인 최적 궤적을 만족하는 특정 값에 수렴하여 유지하는 결과를 보여준다.

4. 결론

본 논문은 벨만 최적 방정식을 등식 제약 조건으로 구성한 제약 최적화 문제 프레임워크를 제안하였으며, 이를 기반으로 KKT 조건에 수렴하는 온라인 강화학습 기반의 최적 제어 기술을 제안하였다. 제안된 프레임워크는 기존의 학습기반 제어 방법에서 발생하는 학습 과정에서의 불안정성 발생에 대한 문제를 개선하였으며, 최적해에 대한 수렴과 학습 과정에서의 성능을 모바일 로봇 어플리케이션의 실험을 통해 검증하였다.

향후 연구에서는 제안된 프레임워크의 예측 기반 미래 제약 조건을 고려하여 장애물이 존재하는 복잡한 환경에서의 학습 문제로의 확장과 학습 과정의 안정성 분석을 통한 이론적 보강을 추가할 계획이다.

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

참고문헌

[1] Frank. L. Lewis, W. E. Dixon, and et al, “A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems” *Automatica,*, vol. 49, no. 1, pp. 82–92, Jan. 2013.

[2] D. Bertsekas “Reinforcement Learning and Optimal Control” *Athena Scientific,* 2019.

[3] R. S. Sutton and A. G. Bartro, “Reinforcement Learning: An Introduction” *The MIT Press,* 2018.

[4] Frank. L. Lewis and et al, “Optimal and Autonomous Control Using Reinforcement Learning: A Survey,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, June. 2028