Check for updates

Journal of Korea Robotics Society (2022) 17(1):076-085
https://doi.org/10.7746/jkros.2022.17.1.076 ISSN: 1975-6291 / eISSN: 2287-3961

제어 장벽함수를 이용한 안전한 행동 영역 탐색과 제어 매개변수의 실시간 적응

Online Adaptation of Control Parameters with Safe Exploration by Control Barrier Function

김 수 영[1] · 손 흥 선†
Suyeong Kim[1], Hungsun Son†

Abstract: One of the most fundamental challenges when designing controllers for dynamic systems is the adjustment of controller parameters. Usually the system model is used to get the initial controller, but eventually the controller parameters must be manually adjusted in the real system to achieve the best performance. To avoid this manual tuning step, data-driven methods such as machine learning were used. Recently, reinforcement learning became one alternative of this problem to be considered as an agent learns policies in large state space with trial-and-error Markov Decision Process (MDP) which is widely used in the field of robotics. However, on initial training step, as an agent tries to explore to the new state space with random action and acts directly on the controller parameters in real systems, MDP can lead the system safety-critical system failures. Therefore, the issue of 'safe exploration' became important. In this paper we meet 'safe exploration' condition with Control Barrier Function (CBF) which converts direct constraints on the state space to the implicit constraint of the control inputs. Given an initial low-performance controller, it automatically optimizes the parameters of the control law while ensuring safety by the CBF so that the agent can learn how to predict and control unknown and often stochastic environments. Simulation results on a quadrotor UAV indicate that the proposed method can safely optimize controller parameters quickly and automatically.

Keywords: Automatic Gain Tuning, Reinforcement Learning, Control Barrier Function, And Safe Exploration

1. 서 론

시스템의 제어기를 설계하기 위해서는 시스템을 이해하고 제어인자들을 최적화해야 한다. 하지만, 대부분의 시스템에는 마찰, 점성 항력, 알 수 없는 토크 및 기타 역학과 같이 정확하게 측정하기 어려운 불확실성이 있으며, 이는 시간이 지남에 따라 지속적으로 변경된다. 또한 기기 마모나 시스템 손상 등 고장이나 오래된 부품을 새 구성 요소로 교체하는 것과 같은 갑작스러운 변경이 있을 수 있다. 이런 이유로 인해 제어 엔지니어는 시스템의 정확한 모델링을 하기에는 제한적이다. 특히 최신 제어 기술은 제어기를 설계하기 전에 시스템을 수학적으로 특성화해야 하는 수학적 모델에 의존하며 물리적 시스템의 수학적 모델이 먼저 생성된 다음 제어기가 모델에 맞게 설계되고 제어기가 물리적 시스템에 구현된다. 이는 모델과 실제 시스템 간에 상당한 차이가 있는 경우 제어기가 제대로 작동하지 않고 시스템의 불안정성을 유발할 수 있는 문제를 가지고 있다. 또한 수학적 모델을 정밀하게 모델링하더라도 이러한 시스템 불확실성 때문에 일반적으로 대부분의 접근 방식은 집약적인 시뮬레이션을 기반으로 경험적으로 개발된다. 이러한 경험적 제어 프로세스는 일반적으로 시스템이 안정적이어야 하기 때문에 제어 파라미터의 안정화에 대한 정확한

Received : Jan. 3. 2022; Revised : Feb. 11. 2022; Accepted : Feb. 16. 2022
※ This project was partially funded by Development of Drone System for Ship and Marine Mission (2.200021.01) of Institute of Civil-Military Technology Cooperation, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2020R1F1A1075857), respectively
1. Masters Student, Dept. of Mechanical Engineering, Ulsan National Institute of Science and Technology, Ulsan, Korea (ksy204104@unist.ac.kr)
† Professor, Corresponding author: Dept. of Mechanical Engineering, Ulsan National Institute of Science and Technology, Ulsan, Korea (hson@unist.ac.kr)

Copyright ⓒKROS

정보가 특정되어야 하며, 원하는 성능을 얻기 위해서 수동 튜닝이 요구된다. 이는 많은 시간과 노력이 소요되며 제어 매개변수의 효과를 정확히 이해하고 있는 제어 전문가에 의해서 조정된다. 따라서 제어 파라미터의 안정화에 대한 기준 확보, 안전한 자동 제어 매개변수 조정을 위한 자동화 제어 기술이 발전하였다.

데이터 기반의 강화학습 제어기는 시스템에 대한 가정이 거의 없어 적절한 데이터의 구축과 보상함수의 설계에 의해 다양한 시스템에 적용될 수 있는 유연성을 가지고 있어 각광받고 있다. 강화 학습은 시행착오 학습을 통해 우수한 제어 기능을 개발하여 제어기가 우수한 제어 전략에 적응할 수 있도록 충분한 유연성을 갖출 수 있다. 또한, 학습자의 행동이 합리적으로 좋은 제어 전략에 정착하면 강력한 제어 기술을 사용하여 기존의 수학적 모델 기반 제어기보다 비교적 안정적이고 더 좋은 성능을 보인다. 하지만 이러한 유연성은 시스템에 대한 배경지식이 전무하다는 가정에서 비롯된 것이므로, 학습 중 시스템의 안정성을 해치는 바람직하지 않은 제어 전략으로 이어질 수 있다. 일반적인 강화학습 시뮬레이션과 실제 환경에서, 상대적으로 단순한 시스템에 대한 안정적인 제어전략을 찾기 위해서도 기본적으로 수백만 건의 불안정한 제어 전략이 나타나는 것을 확인할 수 있고, 이는 시스템의 제어와 운영의 실패로 생각될 수 있다. 엔지니어는 이러한 제어전략이 불안정한 시스템의 동작으로 이어질 수 있다는 점에 유의하는 제어 전략의 영역을 강력하게 제한하는 제약조건을 도입하는 것이 중요하다. 이러한 방법의 예시로 기본 수학적 모델을 기반으로 시스템 모델링의 불확실성을 보정하는 것이 목적인 모델 기반 강화학습 방법, 학습자의 행동 영역을 리아프노프 안정성 이론에 기반하여 제한하는 제약 최적화 기법 등이 사용된다.

본 연구의 목적은 강화학습 기반의 제어기를 사용하여 시스템의 불확정성에 적응할 수 있는 모델의 충분한 유연성을 갖춤과 동시에 안전성 조건에 의해 제어 입력 공간을 제약함으로써 안전한 제어전략을 갖춘 프레임워크를 구축하는 것이다. 강화 학습 네트워크를 훈련하는 동안 PID 제어 매개변수 튜닝 문제에 대한 안전성을 고려한 [1]과 [2]의 작업이 참조되었으며 드론 시뮬레이션에서 제어기의 목표경로 추적 성능실험과 외란에 대한 강건성 실험을 통해 제시된 연구들과 비교함으로써 개선된 점을 보여준다.

2. 선행 연구 조사

일반적으로 자동 제어기 튜닝은 주어진 성능 측정을 최적화하는 제어기 매개변수를 찾는 것을 목표로 한다. 그러나 이러한 매개변수와 성능 값 간의 비선형 매핑은 선험적으로 알려지지 않았으며 복잡한 제어 전략을 예측하는 방법으로 강화학습이 떠오르고 있다. 강화 학습은 학습자의 시행착오를 통해 환경(제어 시스템)과 상호 작용하고 보상과 처벌을 통해 행동(제어 매개변수)을 수정함으로써 제어 매개변수와 성능 간의 매핑을 모델링한다[3]. 강화학습 기반 제어 기술은 과거부터 적응형 PID 제어 기술과 결합되어 강력한 잠재력을 입증하는 많은 연구들이 제시되었다[4-8]. 이러한 연구들은 환경과 시스템에 특성에 따른 적응 제어 매개변수 조정에서 좋은 성능을 보였지만, 강화학습 알고리즘에 효과적으로 제약조건을 적용하여 안전하게 학습자를 훈련시키는 알고리즘은 그 구현의 어려움과 복잡성으로 인해 제시되지 못하였다. 연구 [9]에서는 라그랑주 승수법을 사용하여 강화학습 행동자에 1차원 제약조건을 사용하는 연구가 제시되었다. 어플리케이션으로 4족보행 로봇 관절의 각도 제어에 적용되어 기존의 강화학습 제어기에 비해 성공적인 제어 전략[10]을 제시하였다. 하지만, 제시된 알고리즘을 사용한 인공신경망의 행동자 최적해는 인공신경망 매개변수의 야코비안(jacobian) 행렬과 헤시안(hessian) 행렬의 역행렬을 요구하고 이는 매개변수의 수가 상대적으로 많은 인공신경망 기반 알고리즘에 계산 비용의 증가와 및 알고리즘의 복잡함을 야기하였다. 라그랑주 승수법의 backward 연산을 피하기 위해 로봇 액추에이터의 PWM, RPM 등 저 수준 상태변수를 분석하여 안전한 제어전략으로 제약하는 강화학습 알고리즘에 대한 연구[2]도 제시되었지만 이러한 저수준 상태변수의 제약은 정확한 수학적 시스템 모델링 기반하며 실험적으로 분석되어야 하며, 시스템 모델링 비용을 줄일 수 있는 데이터 기반 알고리즘의 장점을 희석시킬 수 있다.

또 다른 방법은 미지의 함수 공간의 규칙성 가정에 기반한 베이지안 최적화 방법이다. 최근의 연구는 베이지안 최적화 함수의 출력 필드를 최소화하여 소수의 매개변수 조합에 대해서만 함수를 평가하고 전역 최적성을 빠르게 증명하는 실용적인 최적화 알고리즘을 제시하며[11,12], 성능 함수 평가에 대한 불확실성을 명시적으로 모델링하여 실제 시스템에 대한 유연한 예측이 가능하다. 제약조건을 고려한 베이지안 최적화 알고리즘인 SafeOpt[12]는 가우시안 과정 기반 베이지안 최적화 알고리즘을 이용하여 제어 매개변수와 함수의 성능의 전역적 관계를 파악하고, 안전을 보장하는 제어 매개변수에 대해서만 평가하여 학습 시 발생하는 안전 위반 문제를 극복하였다. 이 알고리즘을 응용하여 드론의 PID 제어 매개변수를 조정하는 연구[1,13], 예측제어기법에 응용하는 연구[14], 강화학습의 마르코프 결정 과정에 응용하는 연구[15]들이 적용되었다. 그러나 이 방법은 알려지지 않은 함수 상태 공간에 대해 안전한 예측을 위해서는 시스템의 안전함이 보장된 초기값과 시스템 모델링 불확실성의 정확한 사양을 필요로 한다. 따라서 초기 제어 매개변수 값과 시스템의 불확실성에 대한 보수적인 가정이 요구된다.

반면 리아프노프 함수는 시스템 불확실성에 대한 실험적 탐색과 정확한 해의 계산이 필요하지 않고 시스템을 평형상태로 수렴시키는 속성을 확인하여 상태변수 집합의 안정성을 증명할 수 있으며[16], 더 나아가 리아프노프 안정성 이론에 기반하여 고 수준 상태변수의 제약만으로 저 수준의 제어입력을 제약할 수 있는 제어장벽이론에 대한 연구가 제시되고 있다. 제어장벽이론은 집합 내의 상태변수를 한 점에 수렴시키는 것과 달리 집합의 순방향 불변성을 증명하는 것이 목적으로[17], 집합의 경계에 다다르면 제어장벽함수의 값이 발산하여 학습자에게 강력한 패널티를 주며, 학습자의 행동공간을 제약한다. 알고리즘의 단순함과 강력함으로 인해 다양한 최적 제어[18-20]및 적응형 항법 제어[21-24] 등에 사용되었다.

본 논문에서도 제어장벽함수이론을 강화학습의 기법에 적용하여 인공신경망 훈련 중 시스템의 안정성을 보장하기 위한 문제에 해법을 제시한다. 특히, 기존 시스템에 존재하는 PID 제어기를 효과적으로 활용하여 제약 최적화의 backward 연산을 피했고, 드론의 시뮬레이션에 적용하여 드론의 빠른 제어 속도에 맞는 실시간 계산을 가능하게 하였으며, 제어 매개변수를 환경에 맞게 조정하는 적응형 알고리즘을 제시하였다.

3. 기반 연구

3.1 강화학습

제어 매개변수와 성능평가의 비선형 관계를 예측하기 위한 강화학습 기반의 문제는 현재 상태 $\mathbf{x}$ 에서 보상 R로 이루어진 가장 높은 기대 수익 $G=\sum_{i=0}^{\infty}\gamma^i R_i$을 얻기 위한 최적의 행동 a을 찾는 것이다. 여기서 γ는 0과 1사이에 있는 할인 계수로 미래 단계에 대한 보상의 영향을 약화시키는 역할을 한다. 하지만 현재 단계에서 미래 단계의 보상은 정해지지 않았으므로 강화학습 문제는 이러한 기대수익의 예측 값을 극대화하는 것을 목적으로 한다. 표기의 편의상 본 논문에서는 현재 시간을 나타내는 subscript t를 제거하였다.

환경으로부터 취득한 보상 정보를 이용하는 방법은 두가지로 나뉘는데, 하나는 학습환경 에피소드 끝까지의 보상을 이용한 몬테 카를로(Monte Carlo) 학습, 다른 하나는 시간차(1-step temporal difference) 학습이다.

$$\delta^{TD}=R+\gamma V'-V. \tag{1}$$

여기서 δ^{TD}는 한 단계 시간차(1-step temporal difference) 오차로 다음 상태의 예상 기대수익 $R+\gamma V'$과 현재 상태의 상태가치 V의 차이로 정의된다. 몬테 카를로 학습은 학습자의 학습 궤적에 에피소드의 모든 연속 데이터를 사용하기 때문에 함수 예측의 불확실성이 작다는 장점이 있어 학습에 용의하지만 한 번 학습에 과도한 데이터가 요구된다. 반면 시간차 학습은 한번의 학습에 두개의 데이터가 요구되지만 함수 예측의 불확실성이 매우 커 학습에 용의하지 않은 문제가 있다. 이 두가지 방법의 장점을 고려한 방법이 λ시간차 학습으로, k단계의 시간차 오차를 고려하여 계수 λ만큼 할인한 합 연산 형태로 표현할 수 있다.

$$A=\sum_{i=0}^{k-1}\kappa^i\delta_i^{TD}. \tag{2}$$

여기서 A는 이득으로 정의되며, 강화학습 문제의 목적은 이러한 이득을 극대화하는 것이다.

본 연구에서는 κ시간차 학습기반 연구인 Proximal Policy Optimization[25]을 따라 강화학습 문제가 설계되었다.

$$\begin{aligned}&\max_{\vartheta}\rho(\vartheta)\sum_{i=0}^{k-1}\kappa^i\delta_i^{TD},\\&\text{subject to }\rho(\vartheta)=clip\left(\frac{\pi(a|s)}{\pi_{old}(a|s)},1-\epsilon,1+\epsilon\right).\end{aligned} \tag{3}$$

여기서 ϑ는 인공 신경망의 매개변수, $\rho_t(\vartheta)$는 행동 정책을 이전의 정책과 비교하여 0과 0.2사이의 clipping 상수 ϵ로 제한하여 학습자의 잘못된 행동에 의해 발생한 강한 처벌로 탐색 의지가 저하되는 것을 방지한다. 또한 π는 매개변수화된 행동가 네트워크의 출력인 행동 정책으로, 강화학습 학습자가 취할 행동을 확률적으로 모델링한 확률 분포 함수이며, 정규분포로 가정되어 평균 값이 학습자가 취할 행동, 그 표준편차가 행동의 불확실성이 된다.

3.2 드론 제어 시스템

드론의 제어 문제에서 상태변수 $\mathbf{x}$는 위치 $\mathbf{P}$, 오일러 각 $\boldsymbol{\varsigma}$, 속도 $\mathbf{v}$, 각속도 $\boldsymbol{\omega}$로 이루어진 벡터이고, 1차원 항력 효과를 고려한 시스템 역학 모델을 사용하였다[26].

$$\begin{aligned}\dot{\mathbf{P}}&=\mathbf{v}\\\dot{\boldsymbol{\varsigma}}&=\boldsymbol{\omega}\\\dot{\mathbf{R}}&=\mathbf{R}\hat{\boldsymbol{\omega}}\\\dot{\mathbf{v}}&=-g\mathbf{z}_W+c\mathbf{z}_B-\mathbf{R}\mathbf{D}\mathbf{R}^T\mathbf{v}\\\dot{\boldsymbol{\omega}}&=\mathbf{J}^{-1}\left(\boldsymbol{\tau}-\boldsymbol{\omega}\times\mathbf{J}\boldsymbol{\omega}-\boldsymbol{\tau}_g-\mathbf{A}\mathbf{R}^T\mathbf{v}-\mathbf{B}\boldsymbol{\omega}\right).\end{aligned} \tag{4}$$

여기서 $\mathbf{R}$은 회전행렬, $\hat{\boldsymbol{\omega}}$은 $\boldsymbol{\omega}$의 비대칭행렬, c는 드론의 질량에 의해 정규화 된 추력, 대각행렬 $\mathbf{D}=diag(d_x,d_y,d_z)$는 항력

계수, $\mathbf{J}$는 드론의 관성행렬, $\boldsymbol{\tau}$는 3차원 토크 입력, $\boldsymbol{\tau}_g$는 프로펠러로부터 나오는 자이로스코프 토크, $\mathbf{A}$와 $\mathbf{B}$는 상수행렬이다. 드론의 추력과 토크로 정의된 제어입력 $\mathbf{u}=(c,\tau_x,\tau_y,\tau_z)$는 PID 제어기에서 제어 매개변수 $\mathbf{k}_{pid}$와 목표 상태 $\mathbf{r}$에 의존하며, 제어기 구조 g는 다음과 같이 정의하였다.

$$\mathbf{u}=g(\mathbf{x},\mathbf{r},\mathbf{k}_{pid}). \tag{5}$$

따라서 상태 변수와 제어입력에 의존하는 드론 역학 함수 f는 다음과 같이 도출할 수 있다.

$$\dot{\mathbf{x}}=f(\mathbf{x},\mathbf{u})=f(\mathbf{x},g(\mathbf{x},\mathbf{r},\mathbf{k}_{pid})). \tag{6}$$

여기서 드론의 $\mathbf{k}_{pid}$에 따라서 비행 성능 뿐만 아니라비행 안정성도 크게 달라지게 된다.

3.3 안전 제약조건을 고려한 제어장벽함수

강화학습 알고리즘은 데이터를 이용하여 시스템의 특성에 맞춰 모델링 불확실성을 예측할 수 있고, 이러한 모델 특성에 맞춘 제어전략은 기존의 수학적 모델기반 제어기법에 비해 더 강건한 특성을 가지고 있으나, 훈련 초기에는 시스템에 대한 사전지식이 없기 때문에 안전하지 못한 문제가 있어 실제 어플리케이션에 적용되기에 힘든 점이 있다. 이러한 문제를 해결하기 위해 본 논문은 PID 제어기를 사용하며 시스템의 최소한의 안정성을 보장하는 저 성능 PID 매개변수가 이미 있다는 가정하에 매개변수를 환경에 알맞게 조정하는 문제를 다룬다. 시스템의 불안정성을 가져오는 매개변수가 직접 제어기에 전달되게 하는 것을 방지하기 위해 제어 리아프노프 함수의 지수적 수렴성 조건으로 시스템의 안정성을 먼저 확인하고, 더 안전한 상황에서 PID 매개변수를 업데이트 하기 위해 제어장벽함수를 이용하여 드론의 자세와 로우 레벨 수준의 제어입력을 더 강하게 제한한다. 따라서 제안된 알고리즘은 안정성, 안전성 면에서 저 성능 PID 매개변수와의 비교를 통해 강화학습의 훈련과정에서 안전하게 매개변수를 조정할 수 있다. 제어 리아프노프 함수는 비선형 제어 시스템을 평형점 $\mathbf{x}^*=0$으로 수렴시키는 것을 목적으로 하고 여기서 제어 리아프노프의 지수적 수렴 조건에 여유변수 δ^L를 두어 시스템 안정화 조건을 정의한다[27].

$$L(\mathbf{x})+\lambda\dot{L}(\mathbf{x})\geq\delta^L. \tag{7}$$

상수 λ는 제어 리아프노프 함수의 지수적 수렴 조건을 조정하며 여유변수에 대한 실험적 탐색을 통해 시스템 안정성을 평가할 수 있다.

제어장벽함수는 상태 $\mathbf{x}$가 경계상태 $\bar{\mathbf{x}}$를 벗어나지 못하게 가두어 놓는 역할을 한다. 따라서 벡터 $\mathbf{x}$의 크기가 $\bar{\mathbf{x}}$보다 작아야 하며, 이를 상태 집합 불변성으로 정의한다.

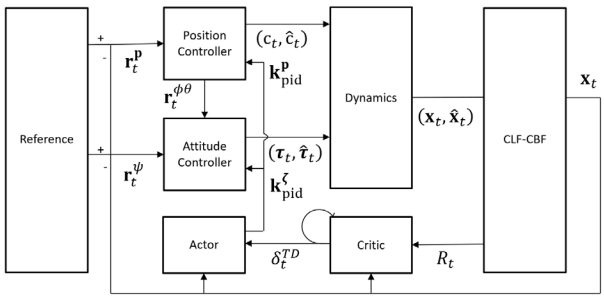
$$|\mathbf{x}|\leq|\bar{\mathbf{x}}|. \tag{8}$$

특히 리아프노프 수렴조건만으로 안전한 제어전략을 추정하기에는 부족한 시스템에 더 강력한 제약조건을 제시하기 위해 제어 리아프노프 함수와 결합되어 사용되었다. 상태변수의 경계상태 제약에 대한 제어장벽함수의 상태 집합 불변성 조건은 다음과 같다[28].

$$B(\mathbf{x},\bar{\mathbf{x}})+\dot{vB}(\mathbf{x},\bar{\mathbf{x}})\geq\delta^B. \tag{9}$$

4. 연구 방법

[Fig. 1]은 강화학습이 적용된 드론의 전체 제어 시스템을 보여준다. 먼저 현재 상태 $\mathbf{x}$에서 강화학습의 행동자 네트워크를 통해 예측된 제어 PID 매개변수 $\hat{\mathbf{k}}_{pid}$를 도출하고, 제어기에는 최소한의 안정성을 보장하기 위해 기본 PID 매개변수가 있어 두가지 매개변수에 대해서 다른 제어 입력 $(\mathbf{u},\hat{\mathbf{u}})$이 출력된다. 각 제어 입력을 고려한 상태 변수 $(\mathbf{x},\hat{\mathbf{x}})$는 시스템 역학을 통해 출력되며 각각의 상태변수는 안정성과 안전성 측면에서 제어 리아프노프 및 제어장벽함수에 의해 평가된다. 안정성 및 안전 조건이 만족되면 추정된 PID 매개변수의 목표 경로에 대한 추적 성능을 평가하여 더 좋은 추적성능을 보이는 제어 매개변수로 업데이트된다. 마지막으로 네트워크를 훈련시키기 위한 마르코프 결정 과정(Markov decision process) 데이터 셋 $(\mathbf{x},\hat{\mathbf{k}}_{pid},\delta^{TD},\gamma)$이 수집된다. 여기서 시간차 오차 δ^{TD}는 평가자 네트워크의 출력인 상태 가치 V와 보상 R를 통해 계산될 수 있고, λ시간차 오차방법을 통해 k단계의 데이터가 모여 PPO 네트워크 훈련에 사용된다.



[Fig. 1] Block diagram of control design

강화학습 문제의 핵심은 제안된 제어 매개변수 $\hat{\mathbf{k}}_{pid}$와 그에 상응하는 학습자의 보상 R의 비선형 관계를 예측하는 것이다. 먼저 PPO 기반의 강화학습 네트워크의 행동자는 드론의 상태변수 $\mathbf{x}$로부터 PID 제어기의 매개변수 벡터를 예측하며, 따라서 행동자의 정책은 평균 $\hat{\mathbf{k}}_{pid}$과 $\boldsymbol{\sigma}_{pid}$의 불확정성을 가진 정규분포를 따른다.

$$\pi_{\vartheta} \sim N(\hat{\mathbf{k}}_{pid}, \boldsymbol{\sigma}_{pid}). \tag{10}$$

강화학습 행동자에 의해 예측된 제어 매개변수는 목적 상태 $\mathbf{r}$와 결합되어 PID 제어기 구조를 통해 제어입력을 출력한다.

$$\hat{\mathbf{u}} = g(\mathbf{x}, \mathbf{r}, \hat{\mathbf{k}}_{pid}). \tag{11}$$

여기서 $\hat{\mathbf{u}}_t$은 강화학습 네트워크에서 출력된 제어 매개변수에 대한 PID 제어기의 출력이며, 이를 이용하여 상태변수를 예측할 수 있고

$$\hat{\mathbf{x}} = \mathbf{x} + f\big(\mathbf{x}, g\big(\mathbf{x}, \mathbf{r}, \hat{\mathbf{k}}_{pid}\big)\big), \tag{12}$$

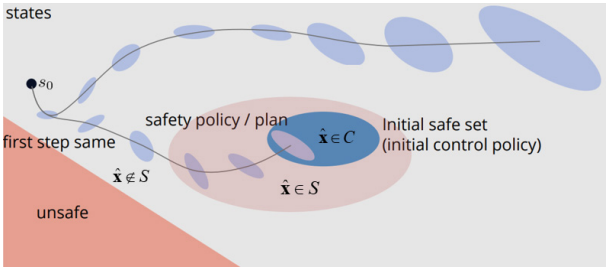
시스템 모델링 불확실성 $\boldsymbol{\epsilon}$을 정의할 수 있다(식 (6)).

$$\boldsymbol{\epsilon} = f\big(\mathbf{x}, g\big(\mathbf{x}, \mathbf{r}, \hat{\mathbf{k}}_{pid}\big)\big) - \dot{\mathbf{x}}. \tag{13}$$

위의 식에서 드론 모델 f와 제어기 g의 구조는 동일하게 고정된 채, 제어변수 입력의 차이에 의해 전체 제어 시스템의 수학적 모델링의 불확실성이 결정된다. 식 (11)과 식 (12)에서 계산된 상태변수와 수학적 모델링의 불확실성은 성능 함수에 포함되어 강화학습의 착취-탐색의 균형 문제[29]에 맞게 고려된다. 특히 이 알고리즘은 시스템의 안정성을 보장하는 PID 매개변수가 존재한다는 가정에서 시작하므로 예측 상태변수가 안정적이지 않을 때는 탐색을 통해 빠른 훈련을 추구하고, 안정상태에 접어들었을 때 착취를 통해 추적 성능을 높이는 상태를 만드는 PID 매개변수를 찾는다.

$$R = \begin{cases} 0.001 \parallel \boldsymbol{\epsilon} \parallel & if\ \hat{\mathbf{x}} \notin S \\ 0.01 \parallel \boldsymbol{\epsilon} \parallel - MSE(\mathbf{r}^{\mathbf{p}} - \hat{\mathbf{p}}) & if\ \hat{\mathbf{x}} \in S \\ 0.5 - MSE(\mathbf{r}^{\mathbf{p}} - \hat{\mathbf{p}}) & if\ \hat{\mathbf{x}} \in C \end{cases}. \tag{14}$$

여기서 $\hat{\mathbf{P}}$은 제안된 제어 매개변수에 의해 예측된 3차원 위치 벡터, $\mathbf{r}^{\mathbf{P}}$는 위치벡터에 대한 목적상태, $MSE(\cdot)$는 제곱 합 평균 연산이며, 집합 S는 시스템의 안정성을 보장하는 상태변수들의 리아프노프 집합, 집합 C는 시스템의 자세제어를 위한 추가적인 제약조건을 만족하는 상태변수들의 제어장벽집합



[Fig. 2] Exploration strategy: A size of cyan ellipse shows the uncertainty of the agent's action on the state space. On unstable region (gray) the agent explores to find the good control strategy. On stable region (pink) exploration and exploitation of the agent's action is balanced. For the stable and safe region (blue), the agent tries to exploit

이다. 만약 PID 매개변수에 의해 도출된 상태변수가 리아프노프 집합에 속해 있지 않을 때($\hat{\mathbf{x}} \notin S$), 학습자는 시스템 불확실성 $\parallel \boldsymbol{\epsilon} \parallel$를 높이는 것을 목표로 하여 빠른 탐색을 추구하며, 시스템 상태가 안정되었을 때($\hat{\mathbf{x}} \in S$), 목표상태에 대한 추적 및 탐색의 균형을 맞춘다. 마지막으로 시스템이 안정적이고 안전하다고 판단되었을 때($\hat{\mathbf{x}} \in C$), 완전히 목표상태를 추적하여 지역 최적점에 갇히지 않고 강화학습 네트워크를 훈련할 수 있다[Fig. 2].

보상구조의 기준은 제어 리아프노프 함수와 제어장벽함수를 계산하여 도출하였다. 제어 리아프노프 함수는 목적 상태에 대한 현재 상태의 볼록함수로 정의할 수 있다.

$$L = \frac{1}{2}(\mathbf{r} - \hat{\mathbf{x}})^T \mathbf{P} (\mathbf{r} - \hat{\mathbf{x}}). \tag{15}$$

여기서 $\mathbf{r}$는 목표 상태 벡터, $\mathbf{P}$는 양의 정부호(positive-definite) 특성을 가지고 있는 대각행렬이다. 제어 리아프노프 함수의 시간에 대한 미분값은 다음과 같으며

$$\dot{L} = \frac{\partial L}{\partial \hat{\mathbf{x}}} \frac{\partial \hat{\mathbf{x}}}{\partial t} = (\mathbf{r} - \hat{\mathbf{x}})^T \mathbf{P} \dot{\hat{\mathbf{x}}}. \tag{16}$$

식 (6)을 통해 상태변수가 안정 집합 S 내부에 존재하도록 하는 안정성 조건을 정의할 수 있다.

$$\hat{\mathbf{x}} \in S \text{ if } \frac{1}{2}(\mathbf{r} - \hat{\mathbf{x}})^T \mathbf{P} \left(\mathbf{r} - \hat{\boldsymbol{x}} + 2\lambda \dot{\hat{x}}\right) \leq \delta^L. \tag{17}$$

본 논문은 제어장벽함수를 이용하여 시스템의 안정성 뿐 아니라 드론의 자세와 저 수준 제어 입력을 제한하는 PID 매개변수를 업데이트하는 알고리즘을 제시한다. 이를 위해 드론의 자세제어에 영향을 미치는 x, y방향 오일러각 $\hat{\phi}, \hat{\theta}$, 제어입력

$\hat{\mathbf{u}}$를 고려하는 새로운 상태변수 $\mathbf{x}^B = (\phi, \theta, \psi)$와 경계상태 $\overline{\mathbf{x}} = (\overline{\phi}, \overline{\theta}, \overline{\mathbf{u}})$를 정의하고, 이를 통해 제어장벽함수 B와 시간에 대해 미분한 함수 $\dot{B}$를 다음과 같이 정의할 수 있다.

$$B = \frac{\{\overline{\mathbf{x}}\}^T \overline{\mathbf{x}} - \{\mathbf{x}^B\}^T \mathbf{x}^B}{\{\overline{\mathbf{x}}\}^T \overline{\mathbf{x}}}, \tag{18}$$

$$\dot{B} = \frac{\partial B}{\partial \mathbf{x}^B} \frac{\partial \mathbf{x}^B}{\partial t} = -2 \frac{\{\mathbf{x}^B\}^T \dot{\mathbf{x}}^B}{\{\overline{\mathbf{x}}\}^T \overline{\mathbf{x}}}.$$

따라서 식 (7)에 의해 상태변수의 안전성 조건을 정의할 수 있다.

$$\hat{\mathbf{x}} \in C \quad if \frac{\{\overline{\mathbf{x}}\}^T \overline{\mathbf{x}} - \{\mathbf{x}^B\}^T (\mathbf{x}^B + 2v\dot{\mathbf{x}}^B)}{\{\overline{\mathbf{x}}\}^T \overline{\mathbf{x}}} \geq \delta^B. \tag{19}$$

이후의 실험에서는 상수 v와 δ^B를 고정시키고 안전 수렴성 조건

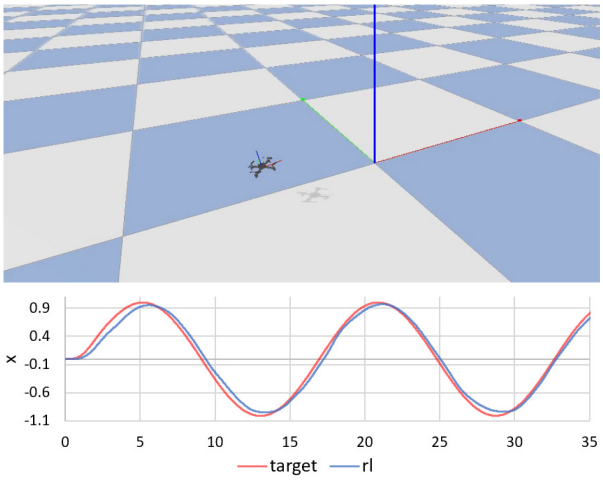
$$B + v\dot{B} = \frac{\{\overline{\mathbf{x}}\}^T \overline{\mathbf{x}} - \{\mathbf{x}^B\}^T (\mathbf{x}^B + 2v\dot{\mathbf{x}}^B)}{\{\overline{\mathbf{x}}\}^T \overline{\mathbf{x}}} \tag{20}$$

의 값을 관찰하며 시스템의 안전함을 확인한다.

5. 연구 결과

5.1 시뮬레이션 환경

시뮬레이션 환경은 드론의 하드웨어와 사양을 구현한 오픈소스 비행 개발 플랫폼 Crazyfile 라이브러리를 사용하였다



[Fig. 3] Quadrotor system simulation environment. Target trajectory is given by a sine wave form in x direction (red) and quadrotor with the proposed adaptive controller (rl) tries to follow the trajectory (blue)

[Fig. 3]. 실험에 사용된 하드웨어 모델은 Crazyfile CF2X로, 무게 27 g와 지름 4 cm의 작은 크기 때문에 빠르고 강력한 제어기가 요구된다. 실제로 시뮬레이션 상에서는 자세 제어에 사용되는 내부 PID 제어기는 200-300 Hz의 주파수로 실행된다. 시스템에 대한 강화학습 문제 정의를 위한 마르코프 결정 과정의 모델링에는 Open-AI Gym 인터페이스가 사용되었고, 강화학습 네트워크 개발용 오픈소스 라이브러리 Baselines를 기반으로 PPO 네트워크를 구현하였다.

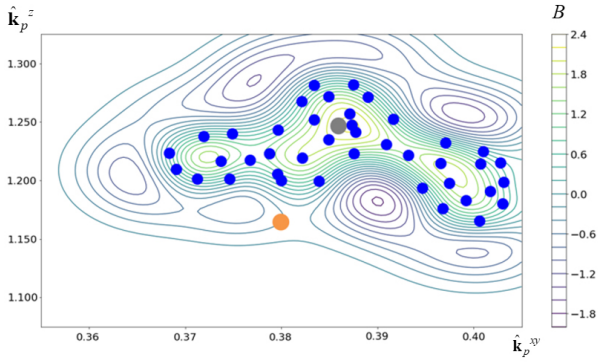
본 논문에서는 두 가지의 비교 시뮬레이션을 진행한다. 첫 번째는 제시된 강화학습 기반 제어기를 일반 PID 제어기, 베이지안 전역 최적화 기법 기반의 제어 매개변수 조정 방법[1]과 비교하여 시뮬레이션 환경에 대한 적응성 및 외란에 대한 강건성을 확인하는 것이 목적이며, 두 번째는 또 다른 강화학습 기반 드론 자세 제어기[2]와 보상의 추이를 비교하여 훈련 중 안전 조건에 대한 만족 여부를 확인한다. 본 연구는 항력, 프로펠러에 의한 지면효과와 같은 외란이 존재하는 환경에서 x방향의 사인파형 목표 경로를 추적한다.

첫 번째 시뮬레이션에서는 제시된 제어 시스템의 환경에 대한 강건성 및 적응성과 추적 성능을 테스트로, 3가지 항목을 분석한다. 첫번째는 제어장벽함수 자체의 값으로 제어장벽함수가 0보다 크면 함수에 정의된 상태변수가 경계상태를 초과하지 않음을 확인할 수 있다(식 (8)). 두 번째는 제어장벽함수의 상태 집합 불변성으로(식 (19)), $B + v\dot{B}$(식 (18))값을 구하여 확인할 수 있다. 세번째는 추적 성능으로, 본 시뮬레이션에서 제시된 목표 경로를 얼마나 잘 추적하는지 위치오차를 이용하여 구할 수 있다.

두 번째 시뮬레이션은 본 연구의 안전성 제약 방법과 다른 시스템의 저 수준 상태변수 직접 제약방법[2]과의 비교로, 적절한 강화학습 학습자의 보상 정도를 정의하고, 이를 통해 훈련 중 안전성 제약 조건의 위반 여부를 판단하고 적절한 훈련을 위해 소모되는 데이터의 수를 파악한다. [2]의 제어기 모델은 동등한 비교 검증을 위해 PPO 인공신경망이 사용되었고, 보상구조도 통일하여 제약조건의 적용 방식의 차이의 효과를 확인하고자 하였다.

5.2 시뮬레이션 결과

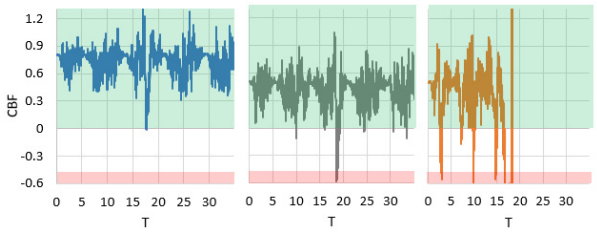
예측된 상태변수가 안전조건 $\hat{\mathbf{x}} \in C$를 만족하기 위해서는 식 (8)의 조건을 만족하여야 하며, 이는 강화학습에서 예측된 제어 매개변수에 대응하는 제어장벽함수 값이 0 이상인 것으로 정의된다. [Fig. 4]는 시뮬레이션 도중 제시된 제어 매개변수가 안전집합 C 내부에서 업데이트되는지 여부를 판단하기 위한 것으로, P 매개변수와 제어장벽함수 값의 상관관계를 보여준다.



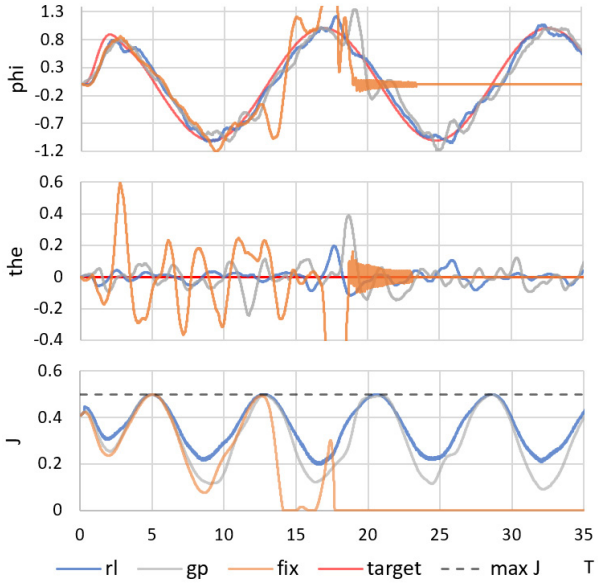
[Fig. 4] Propagation of P gain in safe set: P gain $\hat{\mathbf{k}}_p$ for position x, y is on horizontal axis, P gain for position z is on vertical axis, and the value of the control barrier function B is on z-axis

그래프는 위치 $\mathbf{P}$에 대한 제어장벽함수 B의 등고선을 나타내며 하늘색 등고선($B=0$) 보다 높은 값을 가지는 집합이 안전 집합이다. 그래프 내의 점들은 시뮬레이션에 사용된 각각의 다른 알고리즘에서 사용된 P 매개변수 값으로, 일반적인 PID 제어기에 사용된 제어 매개변수(주황), 가우시안 과정 기반 전역 최적 제어 매개변수(회색), 본 논문에서 제안된 적응형 제어기에서 생성된 제어 매개변수(파랑)를 나타낸다. 시뮬레이션 결과에 따라, 제안된 제어 매개변수의 업데이트가 안전 집합 내에서 잘 이루어진 것을 확인할 수 있으며 강화학습 기반의 제어 알고리즘은 환경 및 상황에 적절한 제어 매개변수를 제시하였다.

제어장벽함수의 값은 현재의 상태가 경계상태집합 내부에 존재하는지 여부를 확인하는데 쓰이지만, 상태집합 불변성의 확인 가능 여부는 알 수 없다. 식 (18)의 조건을 통해 이를 확인할 수 있으며, 본 실험에서는 상수 v를 0.15, δ^B를 0로 정의하였다. [Fig. 5]의 실험은 시뮬레이션 상에서 $B+v\dot{B}$(식 (19))를 관찰하면서 상태변수의 안전 수렴성을 확인한다. 그래프에서는 CBF로 명시 되어있으며, 상태집합불변 조건을 만족하는 구역을 안전 영역, 만족하지 않는 조건을 $B+v\dot{B}$의 값이 -0.5



[Fig. 5] Safety and performance analysis: From top left, the blue-colored graph is a controller with the proposed method, gray-colored graph is a controller based on Bayesian global gain optimization[1] method, and the yellow-colored graph is the low-performance fixed PID controller. Safety is evaluated by slack variable of control barrier function δ^B

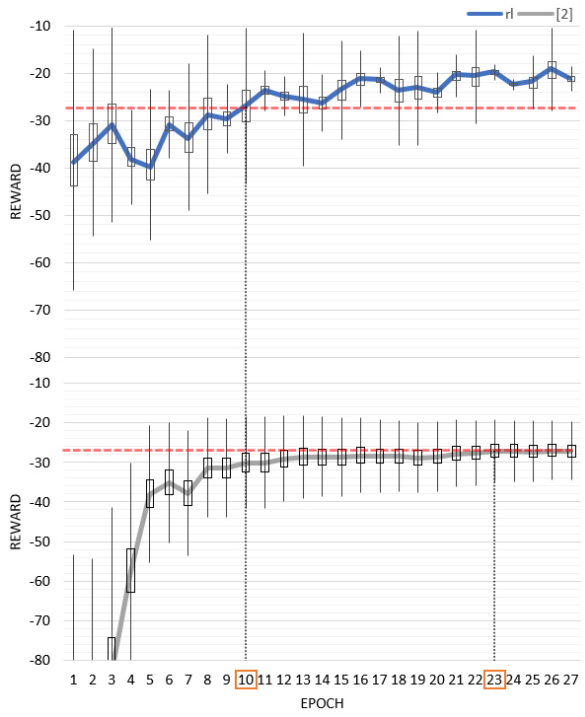


[Fig. 6] Trajectory on the attitude (roll, pitch) in the simulation environment (up). Performance measure J is defined as a negative mean square error of position (down). Given the maximum performance of 0.5 (dashed line)

이하로 정의하여 각각 초록색, 빨간색으로 표시하였다. 실제로 가장오른쪽 그래프의 PID 제어기는 상태변수가 경계상태집합 외부로 발산하였고, 이에 $B+v\dot{B}$값도 발산하였다. 따라서 드론의 안정화에 실패하였고 목표 경로를 추적하지 못하였다. 중간 그래프의 전역최적화 방법은 시뮬레이션 종료까지 목표 경로 추적에는 성공했지만, 제시된 방법에 비해 $B+v\dot{B}$ 값이 전체적으로 낮아 전체적으로 불안정한 제어를 보여주었고, 외란에 대해 상대적으로 강건하지 않은 점을 확인할 수 있다. 마지막으로 가장 왼쪽 그래프의 제안된 강화학습 제어기는 모든 상태변수가 집합 불변 속성을 가졌으며 외란에 대해 강건한 적응형 제어전략을 제시하였다.

이러한 속성은 [Fig. 6]의 사인파형 목표 경로 추적 실험에서도 나타난다. 위쪽 두개의 그래프는 시뮬레이션 경로에 따른 x, y방향의 오일러각을 나타내며, 가장 아래의 그래프는 제시된 목표 경로에 대한 추적 성능을 나타낸다. 이전의 실험과 같이 PID 제어기는 제어에 실패하여 추적에 실패하였고, 전역 최적화 기법은 전반적으로 불안한 자세제어를 보였으며, 제안된 방법이 가장 좋은 자세제어와 추적 성능을 보여주었다.

[Fig. 7]은 이전 실험의 비교군이 아닌 연구 [2]와의 비교로, 강화학습 네트워크 훈련 중 얻은 보상(식 (14))의 평균을 나타낸다. 제안하는 방법은 기존 PID 제어기에 기반한 알고리즘으로 최소한의 안정성을 보장하므로 훈련 초기에도 높은 보상 가치를 얻은 것을 확인할 수 있어 [2]의 연구보다 더욱 안전한 훈련 과정을 제시한다. 보상은 사용된 데이터의 양이 많아질



[Fig. 7] Reward analysis on training epoch. One epoch consists of 2000 episode (14500×2000 = 29 M data points). From top, blue-colored graph is a controller with the proposed method and gray-colored graph is a controller with the low-level state space constraints[2]. And, for the box plot, the box and the bar show the standard deviation within one-sigma and three-sigma

수록 증가하며, 따라서 제어 매개변수 조정의 정도는 보상의 정도와 비례한다고 여겨질 수 있다. 특히, 본 연구에서는 강화학습의 탐색 및 착취문제의 균형을 고려한 보상구조를 제시하였으며, 제안된 방법은 안정적인 제어전략을 제시하는 기준인 보상 -27을 기준으로 10번째 epoch (14500×2000×10 = 0.3B개 데이터)에서 조건을 만족하는 반면 연구 [2]의 드론 자세 제어기는 23번째 epoch (0.6B개 데이터)에서 조건을 충족하며 더 적은 데이터 수를 요구한다. 본 연구의 보상구조에서 학습자는 낮은 좋지 않은 제어 전략에서 불확실성을 높이며 탐색하고, 좋은 제어전략을 찾았을 때 불확실성을 낮추며 착취하며 빠르게 좋은 제어 전략을 탐색할 수 있다.

6. 결 론

본 논문에서는 강화학습을 이용하여 제어 매개변수를 데이터에 맞게 적응적으로 조정하는 방법을 제안하였다. 특히 강화학습 훈련 초기에는 시스템에 대한 기초 정보가 없어 무작위 동작이 제어기에 직접 전달되는 것은 시스템의 안전에 매우 심각한 위험을 끼치기 때문에 안전하게 매개변수를 업데이트하는 방법을 제안하였다. 또한, 제안된 제어기는 베이지안 전역 최적화 방법 기반 제어기에 비해 성능 및 견고성 측면에서 좋은 결과를 보였으며 훈련 중 안전성의 보장 측면에서도 저수준 상태변수 제약 기반의 강화학습 방법보다 더 안전한 제어 전략을 제시하였다.

명 명 법

γ	보상 할인 계수 Discount factor
κ	시간차오차 할인 계수 Temporal difference (TD) error discount factor
k	훈련 시 고려되는 시간차오차의 수 The number of TD error considered on training
ϵ	Clipping 상수 Clipping constant of PPO network
λ	리아프노프 수렴조건 조정 상수 Lyapunov stability constant
υ	상태집합 불변성 조건 조정 상수 Barrier set invariance constant
δ^{TD}	시간차오차 Temporal difference error
δ^{L}	리아프노프 여유 계수 Lyapunov slack variable
δ^{B}	제어장벽 여유 계수 Barrier slack variable
ϑ	인공신경망 매개변수 Neural network parameters
π	행동자 모델 Actor network
$\mathbf{p}$	위치벡터(x, y, z) Position vector
$\mathbf{v}$	속도벡터(v_x, v_y, v_z) Velocity vector
$\boldsymbol{\varsigma}$	오일러각 벡터(ϕ, θ, ψ) Euler angle vector
$\boldsymbol{\omega}$	각속도 벡터$(\omega_x, \omega_y, \omega_z)$ angular veloicy vector
c	추력 scalar thrust
$\boldsymbol{\tau}$	토크(τ_x, τ_y, τ_z) Applied torque on body frame
$\mathbf{u}$	제어입력$(c, \boldsymbol{\tau})$ Control input
$\mathbf{r}$	목표상태변수 벡터$(\mathbf{r}^{\mathbf{p}}, \mathbf{r}^{\mathbf{v}}, \mathbf{r}^{\boldsymbol{\varsigma}}, \mathbf{r}^{\boldsymbol{\omega}})$ Reference state vector
$\hat{\mathbf{k}}_{pid}$	예측 PID 게인 Estimated PID gain
$\boldsymbol{\epsilon}$	시스템 불확실성 System modeling uncertainty

References

[1] F. Berkenkamp, A. P. Schoellig, and A. Krause, "Safe and automatic controller tuning with Gaussian processes," *Workshop on Machine Learning in Planning and Control of Robot Motion, 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, [Online], https://www.dynsyslab.org/wp-content/papercite-data/pdf/berkenkamp-icra16.pdf.

[2] N. O. Lambert, D. S. Drew, J. Yaconelli, R. Calandra, S. Levine, and K. S. J. Pister, "Low level control of a quadrotor with deep model-based reinforcement learning," *arXiv:1901.03737v2 [cs.RO]*, 2019, [Online], https://arxiv.org/pdf/1901.03737.pdf.

[3] F. Berkenkamp, M. Turchetta, A. P. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, Jul., 2017, DOI: 10.1109/TAC.2018.2876389.

[4] A.Y. Zomaya, "Reinforcement Learning to Adaptive Control of Nonlinear Systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 2, Feb., 1994, DOI: 10.1109/21.281435.

[5] X.-S. Wang, Y.-H. Cheng, and W. Sun, "A proposal of adaptive PID controller based on reinforcement learning," *Journal of China Univ. Mining and Technology*, vol. 17, no. 1, 2007, [Online], http://www.paper.edu.cn/scholar/showpdf/MUT2MNzINTD0cx2h.

[6] M. N. Howell and M. C. Best, "On-line PID tuning for engine idle-speed control using continuous action reinforcement learning automata," *Control Engineering Practice*, vol. 8, no. 2, Feb., 2000, DOI: 10.1016/S0967-0661(99)00141-0.

[7] Z. S. Jin, H. C. Li, and H. M. Gao, "An intelligent weld control strategy based on reinforcement learning approach," *The International Journal of Advanced Manufacturing Technology*, Feb., 2019, DOI: 10.1007/s00170-018-2864-2.

[8] M. Sedighizadeh and A. Rezazadeh, "Adaptive PID Controller based on Reinforcement Learning for Wind Turbine Control," *World Academy of Science, Engineering and Technology*, 2008, DOI: 10.5281/zenodo.1057789.

[9] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," *arXiv:1705.10528v1 [cs.LG]*, 2017, [Online], https://arxiv.org/pdf/1705.10528.pdf.

[10] S. Gangapurwala, A. Mitchell, and I. Havoutis, "Guided constrained policy optimization for dynamic quadrupedal robot locomotion," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, Apr., 2020, DOI: 10.1109/LRA.2020.2979656.

[11] F. Berkenkamp, A. P. Schoellig, and A. Krause, "Safe controller optimization for quadrotors with Gaussian processes," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, 2016, DOI: 10.1109/ICRA.2016.7487170.

[12] Y. Sui, A. Gotovos, J. W. Burdick, and A. Krause, "Safe exploration for optimization with Gaussian processes," *32nd International Conference on Machine Learning*, 2015, [Online], http://proceedings.mlr.press/v37/sui15.pdf.

[13] A. K. Akametalu, J. F. Fisac, J. H. Gillula, S. Kaynama, M. N. Zeilinger, and C. J. Tomlin, "Reachability-based safe learning with Gaussian processes," *53rd IEEE Conference on Decision and Control*, Los Angeles, CA, USA, 2014, DOI: 10.1109/CDC.2014.7039601.

[14] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, no. 5, May, 2013, DOI: 10.1016/j.automatica.2013.02.003.

[15] T. M. Moldovan and P. Abbeel, "Safe exploration in Markov decision processes," *arXiv:1205.4810v3 [cs.LG]*, 2012, [Online], https://arxiv.org/pdf/1205.4810.pdf.

[16] A. M. Lyapunov, *The General Problem of the Stability of Motion*, Taylor and Francis Ltd, London, UK, 1992, [Online], https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.910.9566&rep=rep1&type=pdf.

[17] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, Aug., 2017, DOI: 10.1109/TAC.2016.2638961.

[18] K. Galloway, K. Sreenath, A. D. Ames, and J. W. Grizzle, "Torque saturation in bipedal robotic walking through control Lyapunov function-based quadratic programs," *IEEE Access*, vol. 3, pp. 323-332, 2015, DOI: 10.1109/ACCESS.2015.2419630.

[19] A. D. Ames and M. Powell, "Towards the unification of locomotion and manipulation through control Lyapunov functions and quadratic programs," *Control of Cyber-Physical Systems*, vol. 449, 2013, DOI: 10.1007/978-3-319-01159-2_12.

[20] B. J. Morris, M. J. Powell, and A. D. Ames, "Continuity and smoothness properties of nonlinear optimization-based feedback controllers," *2015 54th IEEE Conference on Decision and Control (CDC)*, 2015, DOI: 10.1109/CDC.2015.7402101.

[21] A. Vahidi and A. Eskandarian, "Research advances in intelligent collision avoidance and adaptive cruise control," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 3, pp. 143-153, Sep. 2003. DOI: 10.1109/TITS.2003.821292.

[22] S. Li, K. Li, R. Rajamani, and J. Wang, "Model predictive multi-objective vehicular adaptive cruise control," *IEEE Transactions on Control Systems Technology*, vol. 19, no. 3, pp. 556-566, 2011, DOI: 10.1109/TCST.2010.2049203.

[23] G. J. L. Naus, J. Ploeg, M. J. G. Van de Molengraft, W. P. M. H. Heemels, and M. Steinbuch, "Design and implementation of parameterized adaptive cruise control: An explicit model predictive control approach," *Control Engineering Practice*, vol. 18, no. 8, pp. 882-892, Aug., 2010, DOI: 10.1016/j.conengprac.2010.03.012.

[24] P. A. Ioannou and C. C. Chien, "Autonomous intelligent cruise control," *IEEE Transactions on Vehicular Technology*, vol. 42, no. 4, pp. 657-672, Nov., 1993, DOI: 10.1109/25.260745.

[25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347v2 [cs.LG]*, 2017, [Online], https://arxiv.org/pdf/1707.06347.pdf.

[26] J.-M. Kai, G. Allibert, M.-D. Hua, and T. Hamel, "Nonlinear feedback control of quadrotors exploiting first-order drag effects," *IFAC-PapersOnLine*, Jul., 2017, DOI: 10.1016/j.ifacol.2017.08.1267.

[27] E. D. Sontag, "A Lyapunov-like stabilization of asymptotic controllability," *SIAM Journal of Control and Optimization*, vol. 21, no. 3, 1983, DOI: 10.1137/0321028.

[28] E. Squires, P. Pierpaoli, and M. Egerstedt, "Constructive barrier certificates with applications to fixed-wing aircraft collision avoidance," *2018 IEEE Conference on Control Technology and Applications (CCTA)*, Aug., 2018, DOI: 10.1109/CCTA.2018.8511342.

[29] P. Auer. "Using confidence bounds for exploitation-exploration trade-offs," *The Journal of Machine Learning Research*, vol. 3, pp. 397-422, 2002, [Online], https://www.jmlr.org/papers/volume3/auer02a/auer02a.pdf.

김 수 영

2020 UNIST 기계항공공학과(공학사)

2022 UNIST 기계항공공학과(공학석사)

관심분야: Mechatronics, AI automatic control, navigation, and dynamic system modeling, reinforcement learning

손 흥 선

2000 인하대학교 기계항공공학과(공학사)

2002 Stanford 기계항공공학과(공학석사)

2007 Georgia Institute of Technology 기계공학과(공학박사)

2013~현재 UNIST 교수

관심분야: Mechatronics, sensors and actuators, dynamic system modeling, design optimization, automation, and control