강화학습을 이용한 초고속비행체 제어기 학습

Controller Learning of High-Speed Vehicle using Reinforcement Learning

강민구* 김기응*
MinKu Kang Kee-Eung Kim
* 한국과학기술원
(tominku@kaist.ac.kr)

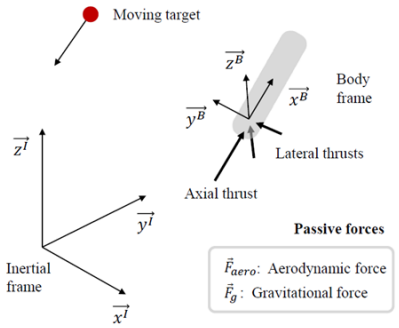
ABSTRACT

비행체가 초고속 비행시 발생하는 양력 (Lift) 및 저항력 (Drag) 과 같은 요소는 시스템에 높은 비선형성을 발생시키는데, 본 연구에서는 이러한 비선형성을 가지는 환경하에서도 데이터 기반 (Data-driven) 국부 최적 (Local optimal) 비행체 제어기 학습이 가능함을 보였다. 본 연구는 데이터 기반 최적 제어 방법론 (강화학습)을 적용함으로써 전통적인 모델 기반 제어 이론적 방법론과 차별화를 달성하였다. 본 연구의 결과물은 개념 증명(Proof of Concept) 수준이지만, 추가적인 hyper-parameter tuning 및 더 많은 컴퓨터 자원 사용을 통해 비행체 제어의 추가적인 성능 향상을 기대할 수 있다.

Key Words : Reinforcement Learning, Reward Function, Nonlinear Dynamical Systems, Controller Learning

1. 서 론

본 연구에서는 아래그림과 같이 중력과 공력 (aerodynamic force)이 존재하는 3 차원 환경하에서 강체로 구성된 비행체가 점으로 표현되는 목표지점까지 접근 하는 상황을 고려한다.



이때 2 개의 좌표계를 고려하는데, 하나는 Cartesian Inertial frame (관성좌표계) 이며 다른 하나는 비행체에 부착된 Body frame (지역좌표계) 이다. 비행체 에이전트 상태의 일부로써 비행체의 위치 (3 차원), 자세 (쿼터니언), 속도, 각속도가 있으며 목표지점의 위치 및 속도는 매 시간간격 마다 측정 (measurement) 이 가능하다고 가정한다. 에이전트의 제어신호는 2 개의 측 추력과 1 개의 길이 방향 축 추력으로 가정한다. 더욱 현실적인 제어신호로써 핀 기반 제어를 생각해 볼 수 있으나, 본 논문에서는 추력기반 제어기만을 고려한다. 본 시뮬레이션 환경은 MuJoCo [1] 물리시뮬레이터 위에서 구성되었으며, 비행체 시뮬레이션 환경 구성에 대한 내용은 본 연구의 사전 연구로써 진행된 바가 있다 [6].

2. 비행체 보상함수 설계

본 연구에서는 비행체에 강화학습 [2] 알고리즘을 적용하기 위하여 아래와 같이 보상함수를 설계 하였다. 보상함수는 포텐셜 함수의 차이 (potential difference) 로 정의하였는데, 본 보상함수를 유한차분법과 체인 룰 (Chain Rule) 을 이용하여 근사하면 최종적으로 미분 항이 들어간 보상함수 G 를 얻게 된다. 강화학습 환경하에서 비행체 에이전트가 학습될 때 매 시간스텝마다 아래 보상함수가 에이전트에게 주어지게 되며 에이전트는 이 보상값의 누적 합을 최대화 하는 방향으로 제어기를 학습하게 된다. 일반적으로 강화학습문제에서 주어진 문제에 대하여 보상함수를 설계하는 것은 매우 어려운 문제로 간주되어진다. 본 연구에서는, 제안한 보상함수가 전통적 비행체 제어기법에서 등장하는 PN Guidance 와 이론적 유사성이 있음을 추가적으로 보였다. 본 연구에서는 기존의 Reward Shaping [4] 형태의 함수꼴 F 로부터 보상함수 유도를 시작한다.

$$F(x_t, a_t, x_{t+1}) = \gamma\phi(x_{t+1}) - \phi(x_t)$$

아래는 위 식으로부터 수정된 본 연구에서 사용된 보상함수꼴을 나타내는데, 포텐셜 차이 (potential difference) 형태로 표현됨으로써 물리적인 의미를 지닐 수 있게 된다.

$$\tilde{F}(x_t, a_t, x_{t+1}) = \phi(x_{t+1}) - \phi(x_t).$$

이 때 포텐셜 함수 (phi) 는 아래와 같이 비행체와 목표지점의 유클리디안 거리에 기반하여 정의하였다.

$$\phi(x_t) := -d(x^{goal},\ x_t)$$
$$d(x^{goal},\ x_t) := \left\| p_t^{target} - p_t \right\|_2$$

또한, 에너지 함수 H 를 포텐셜함수와 크기는 같고 부호만 반대인 함수로 정의하였다.

$$H(x_t) := \left\| p_t^{target} - p_t \right\|_2 = \|L_t\|_2 \quad L_t = p_t^{target} - p_t$$

여기서 기호 L 은 line of sight 를 의미하며, 목표지점과 에이전트의 상대위치 벡터를 나타낸다. 정의된 에너지 함수 (H) 를 위에서 전개한 수식에 대입하고, 유한 차분법 (Finite Difference) 과 체인룰 (Chain Rule)을 이용하여 제안된 보상함수를 정리 (decomposition) 하면 아래와 같다.

$$\tilde{F}(x_t, a_t, x_{t+1}) = -\frac{H(x_{t+1}) - H(x_t)}{\Delta t}\Delta t \approx -\dot{H}(x_t)\Delta t,$$
$$G(x_t, a_t, x_{t+1}) := -\dot{H}(x_t)\Delta t$$

즉, 포텐셜 차이로 표현되었던 보상함수를 아래와 같이 최대 에너지하강 방향 벡터와 상태속도벡터의 내적으로 나타낼 수 있게 된다.

$$\frac{dH(x_t)}{dt} = \left(\frac{\partial H(x_t)}{\partial x_t}\right)^T \frac{\partial x_t}{\partial t} = \left(\nabla_x H(x_t)\right)^T \dot{x}_t$$
$$G(x_t, a_t, x_{t+1}) = -\left(\nabla_x H(x_t)\right)^T \dot{x}_t\ \Delta t.$$

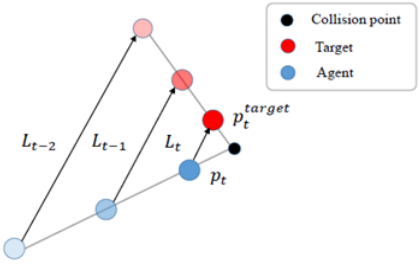
이 때, 에너지함수는 line of sight 벡터만의 함수로 정의하

였으므로 아래와 같이 나타낼 수 있다.

$$\frac{dH(x_t)}{dt} = \left(\frac{\partial H(x_t)}{\partial x_t}\right)^T \frac{\partial x_t}{\partial t}$$
$$= \left(\frac{\partial H(L_t)}{\partial L_t}\right)^T \frac{\partial L_t}{\partial t} = \left(\nabla_L H(L_t)\right)^T \dot{L}_t.$$

$$G(x_t, a_t, x_{t+1}) = inner\left(\frac{1}{\|L_t\|}L_t,\ \left(\dot{p}_t - \dot{p}_t^{target}\right)\right)\Delta t$$

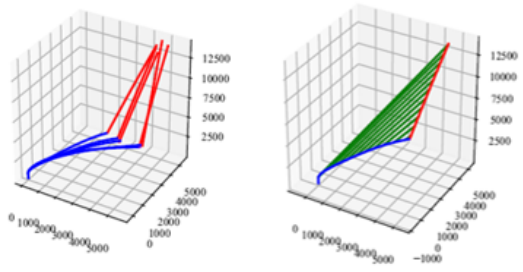
즉, 본 연구에서 제안된 보상함수는 위와 같이 두 벡터의 내적형태로 표현이 가능하다. 이 때 첫 번째 벡터는 normalized line of sight 를 의미하고 두 번째 벡터는 목표지점과 에이전트의 상대속도 벡터를 의미한다. 유도된 결과에 따르면, 본 보상함수를 받으며 학습되는 에이전트는, 에너지필드 H 의 급강하 방향과 상대속도 벡터가 잘 정렬되는 방향으로 제어기를 학습하게 된다고 볼 수 있다. 이를 그림으로 나타내면 아래와 같다.



에이전트와 목표지점의 상대속도벡터가 항상 line of sight 와 정렬되어 있다면 위 그림과 같이 충돌 삼각형 (collision triangle)을 형성할 것이다. 이 충돌 삼각형은 전통적 비행체 제어기법 [5] 에서도 등장하는 개념이며, 이를 통해 본 연구에서 제안된 data-driven 제어기 학습 방법이 모델기반 제어기 학습 방법과 일맥상통하는 측면이 존재한다는 것을 보여준다.

3. 실험결과

비행체 에이전트를 강화학습 알고리즘중 하나인 PPO [3] 를 사용하여 학습한 결과를 아래 그래프에 나타내었다. 목표지점의 초기 위치는 매 episode 마다 random 하게 생성되었다.



이렇게 목표 궤적의 무작위성이 존재하는 상황에서도 학습된 비행체 에이전트가 목표지점에 의도한 바와 같이 접근하는 것을 확인할 수 있었다. 이 때 그래프 축에 나타난 수치는 m 단위 이며 빨간 선은 목표지점의 궤적을 나타낸다. 오른쪽 그래프는 매 시간 간격마다 line of sight 를 궤적과 함께 plot 한 결과를 보여 준다. 보상함수를 유도하는 과정에서 나타난 바와 같이 목표지점과 비행체의 line of sight 가 시간이 변함에 따라 비교적 일정하게 유지 되는 것을 확인할 수 있다.

4. 결론

본 연구에서는 높은 비선형성이 존재하는 시뮬레이션 환경하에서 데이터 기반의 비행체 제어기 학습이 가능함을 보였다. 본 방법론은 기존 고전적 방법론에서 사용하는 비선형 모델의 선형화 (Linearization) 와 같은 근사 (Approximation) 과정을 필요로 하지 않는다. 강화학습과 같은 데이터 기반 제어기법에서는 보상함수 (또는 비용함수) 설계가 결과 최적제어기에 많은 영향을 미치게 되는데, 본 연구에서는 비행체 학습에 적합한 보상함수를 설계하였고 제안된 보상함수를 통해 데이터 기반의 비행체 학습이 가능함을 보였다. 또한 제안된 데이터 기반 방법과 전통적 모델 기반 PN Guidance [5] 와의 연관성을 보였다.

참고문헌

[1] Todorov, Emanuel, Tom Erez, and Yuval Tassa. ”Mujoco: A physics engine for model-based control.” Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on. IEEE, 2012.

[2] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. Vol. 1. No. 1. Cambridge: MIT press, 1998.

[3] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).

[4] Ng, Andrew Y., Daishi Harada, and Stuart Russell. "Policy invariance under reward transformations: Theory and application to reward shaping." ICML. Vol. 99. 1999.

[5] Siouris, George M. Missile guidance and control systems. Springer Science & Business Media, 2004.

[6] MinKu Kang and Kee-Eung Kim, A Simple Physics-based Aerial Vehicle Agent, ICCAS 2018

후기

This work was conducted at High-Speed Vehicle Research Center of KAIST with the support of Defense Acquisition Program Administration (DAPA) and Agency for Defense Development (ADD).