Journal of Korea Robotics Society (2022) 17(1):001-007
https://doi.org/10.7746/jkros.2022.17.1.001 ISSN: 1975-6291 / eISSN: 2287-3961

노이즈 환경에서 효과적인 로봇 강화 학습의 정책 탐색 방법

Effective Policy Search Method for Robot Reinforcement Learning with Noisy Reward

양 영 하[1] · 이 철 수†
Young-Ha Yang[1], Cheol-Soo Lee†

Abstract: Robots are widely used in industries and services. Traditional robots have been used to perform repetitive tasks in a fixed environment, and it is very difficult to solve a problem in which the physical interaction of the surrounding environment or other objects is complicated with the existing control method. Reinforcement learning has been actively studied as a method of machine learning to solve such problems, and provides answers to problems that robots have not solved in the conventional way. Studies on the learning of all physical robots are commonly affected by noise. Complex noises, such as control errors of robots, limitations in performance of measurement equipment, and complexity of physical interactions with surrounding environments and objects, can act as factors that degrade learning. A learning method that works well in a virtual environment may not very effective in a real robot. Therefore, this paper proposes a weighted sum method and a linear regression method as an effective and accurate learning method in a noisy environment. In addition, the bottle flipping was trained on a robot and compared with the existing learning method, the validity of the proposed method was verified.

Keywords: Robotic Arm, Reinforcement Learning, Bottle Flipping, Policy Search

1. 서 론

로봇은 각종 산업 현장과 서비스 분야에서 활발하게 사용되고 있다. 일반적인 로봇은 고정된 환경에서 반복적인 작업을 수행한다. 전통적인 로봇 제어 방식으로는 환경 및 사물의 변화가 복잡한 문제를 해결하기 어렵다. 로봇 주변의 환경과 사물 등 로봇과의 상호 작용의 물리적 모델을 만들지 못하면 로봇의 동작 계획이 불가능하기 때문이다.

로봇 주변의 환경과 사물의 물리적 모델을 해석하기 힘든 상황에서 강화 학습을 적용하여 문제를 해결할 수 있다. 일반적으로 로봇의 강화 학습은 로봇 주변의 환경과, 사물의 상호 작용 동을 측정 및 계산하지 않으며 로봇이 동작을 수행하는 도중 또는 수행한 후 그 결과를 점수로 환산하여 더 높은 점수가 나올 확률이 높은 동작으로 개선한다. 주어진 목적 달성에 가까울수록 높은 점수를 도출하기 때문에 로봇과 환경, 사물의 거동을 정확하게 측정 및 계산할 수 없어도 목적을 달성할 수 있다.

지금까지 강화 학습으로 로봇과 같은 에이전트에게 특정 일을 학습시키는 연구가 다양하게 이루어졌다. 4족 보행 로봇의 보행운동[1], 로봇 팔의 중량 들어올리기나[2] Ball-in-a-cup 놀이[3], 휴머노이드 로봇의 활쏘기[4] 등 로봇 분야는 물론 자율주행 차량의 장애물 회피 기동[5]과 같이 에이전트가 스스로 학습하는 어떤 것이라도 강화 학습을 적용할 수 있으며 이에 대한 연구가 많이 이루어지고 있다[6,7].

많은 강화 학습의 연구가 가상의 환경이나 매우 정교하게 만들어진 로봇과 환경에서 이루어진다. 하지만 실제 로봇의 강화 학습에 많은 제약이 따른다. 모델이 부정확할 수 있고, 시도 횟수를 제한없이 많이 늘릴 수도 없다[8]. 또 노이즈가 많은 환경에서 학습을 할 수도 있다. 그래서 효율적이고 빠른 학습 방법에 대하여 보상 함수, 정책 탐색 방법 등 다양한 방면으로

Received : Jul. 20. 2020; Revised : Jan. 18. 2022; Accepted : Feb. 10. 2022
※ This project was funded by Sogang University Research & Business Development Foundation
1. MS Student, Mechanical Engineering, Sogang University, Seoul, Korea (dakmuk@sogang.ac.kr)
† Professor, Corresponding author: Mechanical Engineering, Sogang University, Seoul, Korea (cscam@sogang.ac.kr)

연구되었다[9-16].

실제 로봇의 학습의 경우 보상의 노이즈로 학습이 실패하는 경우가 있다. 로봇의 반복성이 완벽하지 않기 때문에 보상은 확률 분포를 나타내 같은 동작 변수에도 보상은 계속 달라질 수 있다. 또 높은 보상을 기대할 수 있는 움직임에서 실제로는 더 낮은 보상이 나올 가능성도 있다. 이러한 학습 과정에서의 노이즈는 로봇이 더 나은 동작을 찾아가는데 방해가 된다. 이전 보틀 플리핑을 로봇에게 학습 시킨 연구에서도 학습 과정에서 보상 값이 감소하며 노이즈로 인해 학습 속도가 감소하는 것을 확인하였다[8].

본 논문에서는 보상의 노이즈가 강화학습의 효율을 저해할 경우 해결할 방법을 제안하였다. 로봇이 더 나은 동작을 찾을 때 동작 변수의 주변을 탐색하는데 이때 탐색하는 점의 수를 늘리면 노이즈의 영향이 줄일 수 있다. 지금 까지의 강화 학습은 학습 속도를 향상시키기 위해서 적은 탐색으로 기울기를 구하는 방법을 발전시켜왔다. 하지만 노이즈가 큰 환경에서는 한번의 탐색 과정에서 더 정확한 방향을 찾는 것이 더 큰 이점으로 작용할 수 있다. 많은 탐색점을 이용하여 방향을 탐색하는 방법으로 보상의 차이를 가중치로 하는 가중 합 방법과 선형 회귀 방법으로 방향을 탐색하는 두 가지의 알고리즘을 제안하였다. 탐색하는 점이 많아지면 동작을 찾아가는데 더 많은 시도가 필요하지만 정확한 동작을 찾아 가기 때문에 결과적으로 학습의 속도는 더 빨라질 수 있다.

이를 검증하기 위해 인간의 어깨, 팔꿈치, 손목을 모사한 3자유도 로봇을 이용하여 보틀 플리핑을 학습했다. 보틀 플리핑은 일부분 액체가 들어있는 병을 공중으로 던져 올린 후 바닥에 착지시켜 수직으로 세우는 놀이이다. 물병 내부의 물의 운동 때문에 물병의 운동을 정밀하게 제어하기는 매우 어려우며 미세한 차이로도 결과는 크게 달라질 수 있다. 같은 동작 변수로 병을 던졌을 때 나타나는 결과는 분산된다. 이와 같이 보상의 정밀도가 낮은 상황에서는 로봇이 더 나은 동작을 찾아가기 어렵다. 본 논문에서는 이러한 노이즈의 영향을 줄이는 학습 방법을 제안하였으며 노이즈를 고려하지 않은 학습 방법과 비교하여 더 빠른 학습이 가능한지 확인해보았다.

실험에 사용된 시스템과 강화 학습의 방법에 대해서는 2장에서 기술하였다. 정책 탐색의 방법에 대해서는 3장에서 기술하였다.

2. 시스템 구성 및 강화 학습

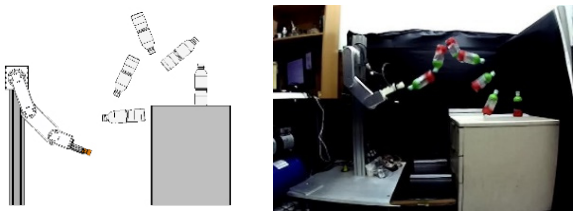
2.1 보틀 플리핑

본 논문에서 로봇에게 학습시킬 것은 [Fig. 1]과 같이물병을 공중에 던져 올린 후 바닥에 수직으로 착지시키는 놀이인 보틀 플리핑이다. 물병안의 물까지 고려하여 물병의 정확한 물리적 모델을 만들기는 대단히 어렵다. 그렇기 때문에 로봇과 사물의 상호 관계를 찾지 않고 목적의 달성만을 고려하는 모델 프리 방식으로 해결한다. 로봇 각 관절의 길이 모터의 입력 펄스 대비 회전각 등의 기구학 및 병의 무게, 액체의 양 등의 환경 조건을 계산하지 않았다.

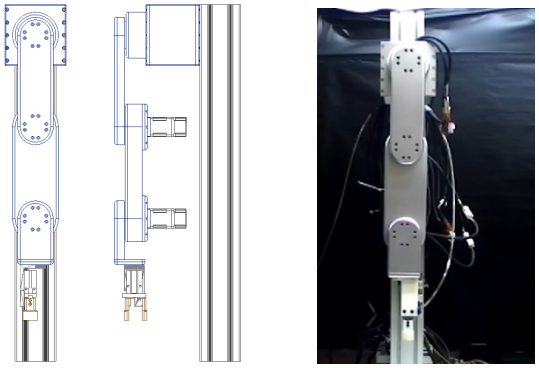
2.2 로봇 시스템 설계

보틀 플리핑을 학습시킬 로봇은 [Fig. 2]와 같이 AC서보 모터와 공압 그리퍼를 이용한 3축 로봇이다. 로봇의 몸체는 3D 프린팅 및 알루미늄 가공으로 제작하였다. [Fig. 2]의 사진의 가장 위쪽의 조인트부터 연결된 순서대로 어깨, 팔꿈치, 손목을 대응시켜 인간의 팔과 유사한 동작을 구현할 수 있도록 하였다.

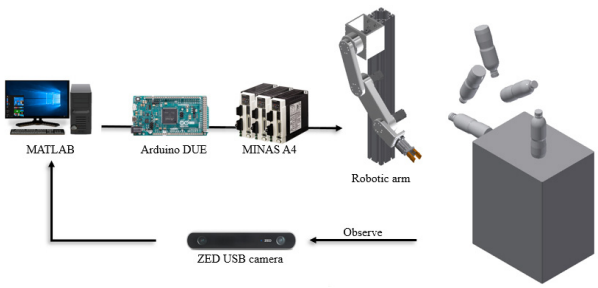
로봇을 제어하는 시스템은 [Fig. 3]와 같이 구성하였다. PC



[Fig. 1] Sketch and sequence photo describing bottle



[Fig. 2] 3DOF Robot Arm



[Fig. 3] System diagram

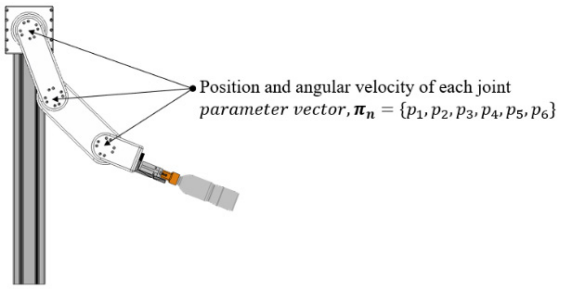
에서 동작 변수를 계산하고 이를 제어보드에서 펄스 신호로 바꾸어 드라이브로 보내 로봇을 제어한다. 로봇의 동작은 USB camera를 통해 초당 100 프레임의 영상으로 촬영되어 강화 학습을 위한 정보로 활용된다. 제어보드는 Arduino DUE를 사용하였고 드라이브와 모터는 Panasonic의 MINAS A4 Series를 사용하였으며 카메라는 ZED stereo camera를 사용하였다.

2.3 동작의 변수화

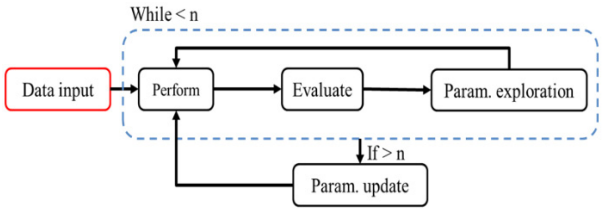
로봇이 동작하는데 많은 변수가 있으면 학습 과정에서 불필요한 계산이 많아진다[1,12]. 본 논문에서는 [Fig. 4]와 같이 물병 던지기 동작을 그리퍼가 물병을 놓는 순간 각 조인트의 각도 및 각속도를 나타내는 6개의 변수로 압축하였다. 물병 던지기는 물병 속의 물의 움직임에도 영향을 받기 때문에 물병을 놓기 전의 움직임이 결과에 영향을 미칠 수 있다. 하지만 이를 모두 고려하는 것은 학습의 복잡성이 커지는데 비해 효과는 적다. 그래서 단순하게 로봇이 병을 놓는 시점에서만 각 조인트의 각도 및 각속도를 고려하였다.

2.4 강화 학습

강화 학습의 과정은 [Fig. 5]과 같다. 먼저 입력 값을 로봇이 받아 동작을 실행하고 보상 값을 구해 동작을 평가한다. 변수 탐색은 입력한 변수에서 작은 변화를 주어 주변에서 보상을 구하는 과정이다. 이를 여러 차례 반복하여 탐색과정에서 구한 정보를 계산하여 최종적으로 변수를 갱신하는 과정을 반복한다.



[Fig. 4] Parameterized motion of bottle throwing



[Fig. 5] General reinforcement learning process

$$\boldsymbol{\pi}_{batch} = \begin{bmatrix} p_{1_{batch}} \\ \vdots \\ p_{i_{batch}} \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_i \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \end{bmatrix} \quad (1)$$

[Fig. 5]의 동작 변수 탐색은 초기 입력 값에 난수를 추가하는 과정으로 식 (1)과 같이 나타낼 수 있다. 난수를 생성하는 범위는 로봇이 동일한 변수에서 동일한 보상을 얻는 반복성에 따라 적당한 크기로 한다. 그 범위가 너무 작으면 높은 확률로 탐색에 악영향을 미칠 수 있다. 변수를 최종적으로 업데이트하기 전까지 탐색하는 과정의 반복을 Batch iteration이라한다.

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k + \alpha \cdot \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} \quad (2)$$

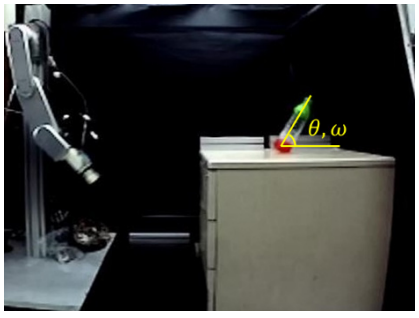
[Fig. 5]의 Param. Update는 Batch iteration에서 사용한 동작 변수와 보상으로 최적의 방향을 구하고 입력 값을 업데이트하는 과정이다. 이를 식 (2)와 같이 나타냈다. α (Step size)는 상수로 매 업데이트마다 동일한 크기를 적용하였다.

2.5 보상 함수

학습의 과정에서 동작을 평가하는 보상 값은 동작의 결과로부터 얻은 정보와 보상 함수를 이용하여 계산한다. 보상 함수는 보틀 플리핑이 성공하는 방향으로 로봇이 학습할 수 있도록 성공에 가까운 동작에 더 높은 보상 값을 얻을 수 있도록 식을 세워야 한다.

$$J = -c_1(a_1 - \theta)^2 - c_2(a_2 - \omega)^2 \quad (3)$$

[Fig. 6]은 동작의 결과를 100FPS의 속도로 촬영하고 물병이 바닥면에 닿는 순간 물병과 바닥면이 이루는 각도 θ와 각속도 ω를 나타낸 것이다. 두가지 정보를 바탕으로 어떤 동작의 결과가 성공에 더 가까운지 식 (3)을 통해 보상을 구하여 평가



[Fig. 6] Angle and Angular velocity measured at moment of landing

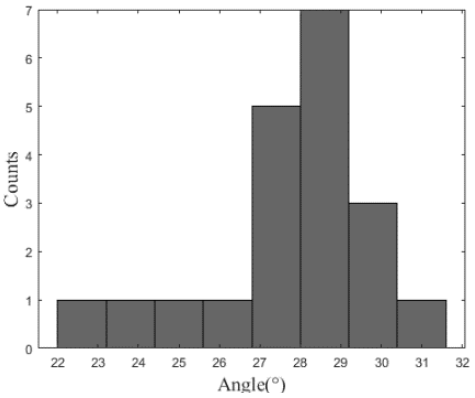
하였다.

식 (3)의 상수 a_1과 a_2는 각각 가장 성공률이 높은 각도와 각속도로 본 논문에서는 물병 던지기 실험을 통해 구한 값으로 a_1은 90 (deg), a_2는 250 (deg/sec)를 사용하였다. 바닥면에 닿는 순간의 각도가 90 deg일 때 보상은 최대가 되고 각속도도 같은 방식으로 250 deg/rad일 때 보상이 최대가 되도록 식을 세웠다. 각도와 각속도는 서로 다른 물리량이므로 이를 합하기 전 상수 c_1과 c_2를 곱하였으며 상수들은 어느 항으로 영향이 치우쳐지지 않도록 적당한 비율로 정하는데 본 논문은 10:1의 비율로 c_1과 c_2를 정하였다.

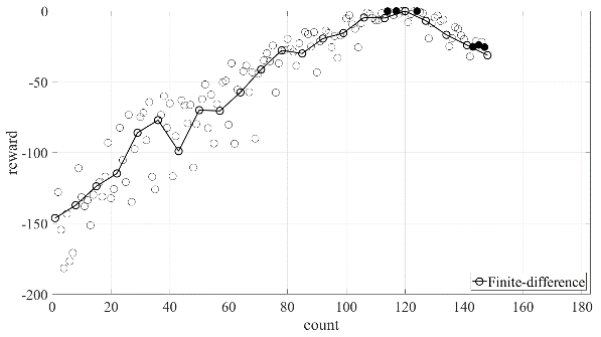
2.6 정책 탐색 방법

$$\boldsymbol{d} = (\boldsymbol{\Delta\Pi}^{\boldsymbol{T}}\boldsymbol{\Delta\Pi})^{-1}\boldsymbol{\Delta\Pi}^{\boldsymbol{T}}\boldsymbol{\Delta\hat{J}} \tag{4}$$

식 (4)는 Finite Difference Method의 정책 탐색 방법을 나타낸 것이다[3]. $\boldsymbol{\Delta\Pi}$는 Batch iteration을 생성할 때 입력 값에 추가한 난수, $\boldsymbol{\Delta\pi_i} \in \mathbb{R}^m$ 벡터를 모은 행렬로 $\boldsymbol{\Delta\Pi} = [\boldsymbol{\Delta\pi_1}, \cdots, \boldsymbol{\Delta\pi_n}]^T$으로 나타낼 수 있다. $\boldsymbol{\Delta\hat{J}}$는 모든 Batch iteration에 대하여 로봇의 임무 수행 후 평가되는 보상 값의 증감, $\boldsymbol{\Delta\hat{J}_n}$를 계산하여 모은 행렬이다. 이는 $\boldsymbol{\Delta\hat{J}} = [\boldsymbol{\Delta\hat{J}_1}, \cdots, \boldsymbol{\Delta\hat{J}_n}]^T$로 나타낼 수 있다. $\boldsymbol{\Delta\Pi}$의 Pseudoinverse와 $\boldsymbol{\Delta\hat{J}}$의 곱을 통하여 구한 기울기 값은 보상 값이 가장 크게 증가할 방향을 나타내며, 이를 식 (2)에 대입하여 변수를 갱신한다. 보상의 노이즈가 없다면 이때 구한 $\boldsymbol{d}$는 최적의 방향을 나타내지만 실제로는 그렇지 않다.



[Fig. 7] Histogram of distributed result of throwing the bottle 20 times



[Fig. 8] Result plot of learning bottle flipping task using finite difference method, filled markers indicate successful cases

로봇이 동작을 반복할 경우 병안의 물의 운동, 그리퍼와 물병사이의 미끄러짐 등의 원인으로 결과는 분산되어 나타난다. 이는 다음 동작을 학습할 때 정확한 방향 탐색에 방해가 된다. [Fig. 7]은 20회 동일한 동작으로 물병을 던져서 바닥에 도달하는 순간의 각도를 히스토그램으로 나타낸 것으로 최대와 최소값의 차이가 약 8~9도 난다. 이러한 노이즈로 인해 갱신 이후 동작으로 오히려 낮은 보상을 얻을 수 있다.

[Fig. 8]은 유한 차분법으로 보틀 플리핑을 학습시킨 결과 보상의 변화를 그래프로 나타낸 것이다. 선으로 연결한 마커들은 변수를 갱신한 다음의 동작에서 보상을 나타낸 것이고 내부가 칠해진 마커는 보틀 플리핑을 성공한 케이스를 나타낸 것이다.

보상이 점점 증가하여 최고점에 도달하고 114번째에 처음으로 성공하는 케이스를 확인할 수 있었다. 다만 학습하는 중간 보상이 낮아지는 경우가 있음을 확인할 수 있다. $\mathbf{d}$를 구하기 위해 유한 차분법으로 Gradient를 계산할 경우 반복성의 한계로 보상에는 노이즈가 포함되고 그로 인해 $\mathbf{d}$는 항상 보상이 증가하는 방향을 보장하지 않는다는 것을 확인할 수 있다.

노이즈를 완화시키는 간단한 방법은 탐색점의 수를 늘리는 것이다. Param. Update 과정까지 탐색의 횟수는 늘지만 매번 정확한 방향을 보장할 수 있다면 결과적으로는 더 빠르게 높은 보상을 기대할 수 있다. 다음 3장에서는 여러 탐색점을 가지고 가중 합과 선형 회귀의 방법으로 방향을 결정하는 방법에 대하여 설명하였다.

3. 노이즈의 영향을 줄이는 정책 탐색 방법

3.1 가중 합을 이용한 방법

탐색 과정에서 각 방향으로 얻는 보상의 증감을 가중치로 탐색점을 합하는 가중 합으로 노이즈를 완화할 수 있다. 이를 식 (5)와 같이 나타냈다.

$$\mathbf{d} = \sum_{i=1}^{n} \Delta\pi_i \Delta\hat{J}_i = \Delta\Pi\Delta\hat{J} \tag{5}$$
$$\boldsymbol{\Delta\Pi} = [\boldsymbol{\Delta\pi_1}, \cdots, \boldsymbol{\Delta\pi_i}]^{\boldsymbol{T}}, \boldsymbol{\Delta\hat{J}} = [\boldsymbol{\Delta\hat{J}_1}, \cdots, \boldsymbol{\Delta\hat{J}_i}]^{\boldsymbol{T}}$$

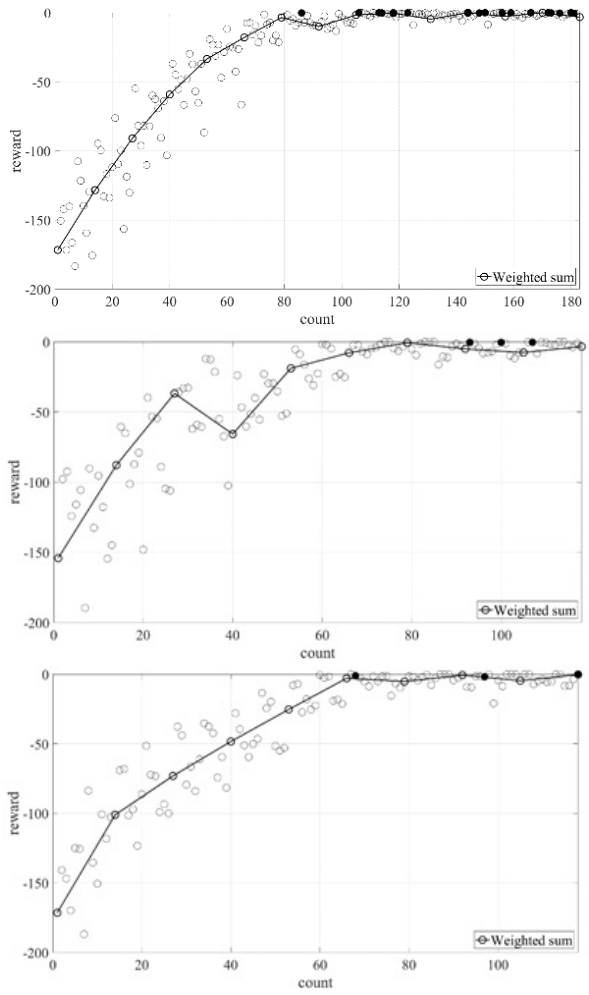
가중 합 방법으로 보틀 플리핑을 학습한 결과는 [Fig. 9]와 같다. 초기 변수는 유한 차분법 실험과 동일하다. 학습을 시작하여 갱신한 것은 선으로 연결하였고 성공한 케이스는 칭해진 마커로 표시하였다. 보상이 최댓값에 수렴할 때까지 변수를 갱신할 때마다 보상이 증가하여 86, 68, 83번째에서 처음으로 성공한 케이스를 확인할 수 있다. 기존 유한 차분법을 이용한 방법에서는 탐색점을 6개를 이용하였는데 가중 합의 방법에서는 12개의 탐색점을 이용하였다. 한번의 탐색에 더 많은 탐색 과정이 필요하여 학습과정이 더 천천히 진행된다. 하지만 더 정확한 방향을 탐색할 수 있었으며 결과적으로는 더 적은 횟수에서 성공시켰다.

3.2 선형 회귀를 이용한 방법

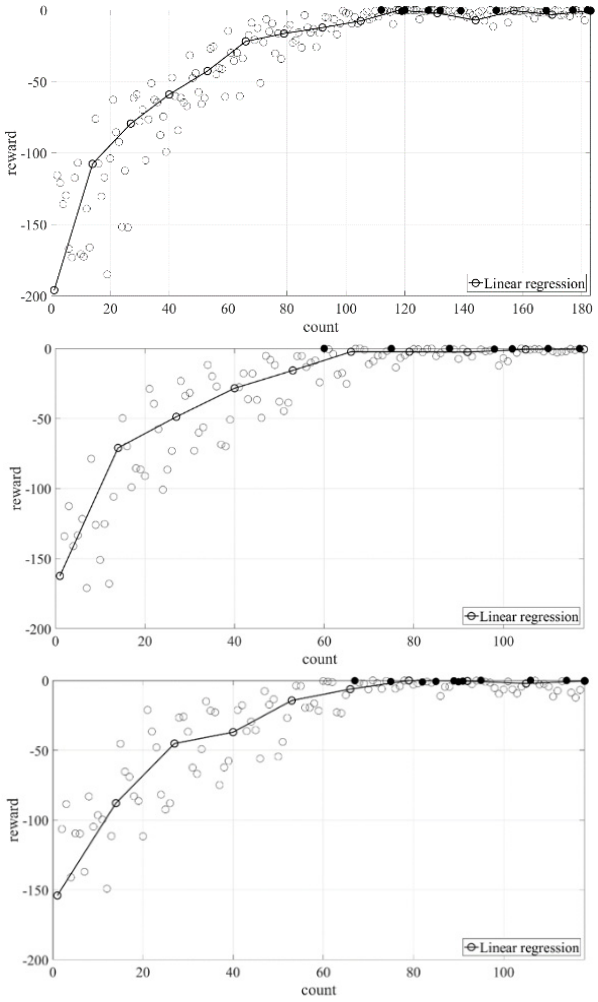
탐색 과정에서 얻는 노이즈가 있는 보상으로부터 실제 Gradient와 가까운 값을 구할 수 있는 방법으로 선형회귀의 방법이 있다. 탐색하는 범위가 좁기 때문에 식 (6)과 같이 선형 근사할 수 있다. 이 방법으로는 많은 탐색점을 사용할 수록 더 정확한 방향을 구할 수 있다. $b_1,\cdots,b_m$ 은 선형 회귀로 근사시킨 식의 각 계수로 물리적 의미는 Gradient와 동일하다. $\boldsymbol{d}=[b_1,\cdots,b_m]$ 는 노이즈의 영향을 완화하여 구한 탐색 방향으로 최적의 방향을 제시할 것이라고 기대할 수 있다.

$$f(p_1,\cdots,p_i)=\sum_{i=1}^{m} b_i p_i + b_0 \tag{6}$$

선형 회귀의 방법으로 보틀 플리핑을 학습한 결과는 [Fig. 10]과 같다. 가중 합 방법과 동일하게 12개의 탐색점을 이용하였다. [Fig. 9]의 가중 합 방법으로 학습한 결과와 마찬가지로 보상이 최댓값에 수렴할 때까지 변수를 갱신하고 보상이 낮아지는 경우는 없었으며 112, 60, 75번째에 처음으로 보틀 플리핑을 성공시켰다.



[Fig. 9] Result plot of learning bottle flipping task using weighted sum method, filled markers indicate successful cases



[Fig. 10] Result plot of learning bottle flipping task using linear regression method, filled markers indicate successful cases

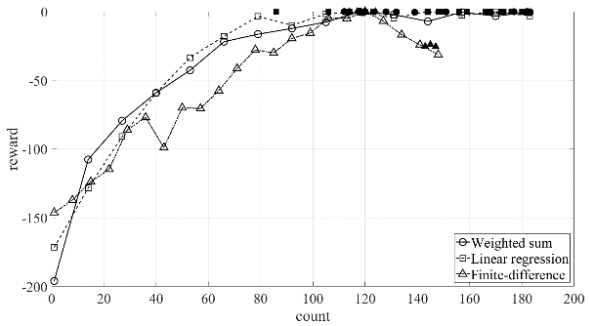
4. 결 론

본 논문에서는 실제 로봇이 강화 학습을 통해 임무를 학습하는 과정에서 얻는 보상의 정밀도가 낮아서 발생하는 탐색의 실패를 줄이는 방법을 제안하였다.

지금까지 많은 로봇의 강화 학습 연구들은 고성능, 고정밀의 로봇위주로 이루어져 왔으며 작업 공간은 매우 정교하게 만들어져 환경으로 인한 노이즈의 영향 잘 고려하지 않았다. 하지만 실제 로봇의 학습은 노이즈가 많은 환경에서 이루어질 수 있다. 그래서 앞으로의 연구를 통해 정밀한 보상을 얻지 못하는 상황에서도 효과적인 학습을 할 수 있는 방법을 찾아야한다.

본 논문에서는 방향 탐색 방법으로 가중 합을 이용한 방법과 선형 회귀를 이용한 방법을 제안하였고 보틀 플리핑을 학습하는 실험을 통해 확인한 결과 두가지 방법 모두 노이즈를 고려하지 않은 방법보다 더 빠른 학습이 가능했다. [Fig. 11]은 각 탐색 방법의 학습 속도를 비교하기 위해 보상의 변화를 한 꺼번에 나타낸 것이다. 가중 합방법은 3회의 실험에서 86, 68, 83번째에 처음으로 보틀 플리핑을 성공시켰고 선형 회귀의 방법은 3회의 실험에서 112, 60, 75번째에 처음으로 보틀 플리핑을 성공시켰다. 이는 유한 차분법을 이용한 방법으로 114번째에 성공했던 것과 비교하여 더 빠르게 성공시킨 것이다.

유한 차분법을 이용한 방향 탐색은 탐색 과정에서 노이즈로 인해 잘못된 방향으로 학습할 가능성이 있으며 이는 보상이 오히려 줄어드는 방향으로 변수가 갱신되기도 하며 잘 수렴하지 않는 모습을 보였다. 한편 가중 합과 선형 회귀 방법으로 방향을 탐색한 경우에는 유한 차분법으로 변수를 갱신할 때와는 달리 보상이 항상 증가하는 방향으로 학습이 잘 되었다.



[Fig. 11] Result plot of learning bottle flipping task, filled markers indicate successful cases

[Table 1] Comparison between finite difference (FDM), weighted sum (WS) and linear regression (LR) method

Method	FDM	WS	LR
Exploration Points	6	12	12
First Successful Case (Avg.)	114	79	82

더 많은 탐색점을 이용하여 방향을 탐색함으로 노이즈의 영향이 줄어들었음을 확인할 수 있었고 또 학습의 효과가 좋아 탐색점의 수가 늘어났음에도 불구하고 더 빨리 임무를 성공시켰다[Table 1]. 하지만 Batch iteration의 크기가 너무 크다면 오히려 보상을 증가시키는데 더 많은 시도 횟수가 필요할 수 있어 비효율적일 수 있기 때문에 적당한 크기로 정해야 할 것이다.

본 논문의 보틀 플리핑 학습에서는 모든 탐색 방법에 대하여 동일한 Step size를 적용하였다. 이는 정확도가 높아진 이점을 완전히 활용하지 않은 것이다. 예를 들면 탐색 방향의 정확도가 향상되었기 때문에 초기 Step size를 키워 더 빠른 학습을 보장할 수 있을 것이다. 후속 연구에서는 정확한 탐색의 이점을 활용하는 방향으로 보완하여 학습의 속도를 극대화할 수 있을 것으로 기대한다.

References

[1] N. Kohl and P. Stone, “Policy Gradient Reinforcement Learning for Fast Quadrupedal Locomotion,” *2004 IEEE International Conference on Robotics & Automation*, New Orleans, LA, USA, pp. 2619-2624, 2004. DOI: 10.1109/ROBOT.2004.1307456.

[2] M. T. Rosenstein and A. G. Barto, “Robot Weightlifting By Direct Policy Search,” *2001 International Joint Conference on Artificial Intelligence*, Seattle, USA, pp. 839-844, 2001, [Online], https://dl.acm.org/doi/abs/10.5555/1642194.1642206.

[3] J. Kober and J. Peters, “Policy Search for Motor Primitives in Robotics,” *Machine Learning*, vol. 84, pp. 171-203, 2011, DOI: 10.1007/s10994-010-5223-6.

[4] P. Kormushev, S. Calinon, R. Saegusa, and G. Metta, “Learning the skill of archery by a humanoid robot iCub,” *2010 10th IEEE-RAS International Conference on Humanoid Robots*, Nashville, TN, USA, pp. 417-423, 2010, DOI: 10.1109/ICHR.2010.5686841.

[5] D. H. Kang, J. H. Bong, J. Park, and S. Park, “Reinforcement Learning Strategy for Automatic Control of Real-time Obstacle Avoidance based on Vehicle Dynamics,” *Journal of Korea Robotics Society*, vol. 12, no. 3, pp. 297-305, Sept., 2017, DOI: 10.7746/jkros.2017.12.3.297.

[6] R. S. Sutton and A. G. Barto, “Introduction,” Reinforcement Learning: An Introduction, 2nd ed. The MIT Press, 2014, ch. 1, sec. 1-7, pp.1-18, [Online], https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf.

[7] M. P. Deisenroth, G. Neumann, and J. Peters, “A Survey on Policy Search for Robotics,” *Foundation and Trends® in Robotics*, vol. 2, no. 1-2, pp. 1-142, 2013, DOI: 10.1561/2300000021.

[8] Y. H. Yang, S. H. Lee, and C. S. Lee, “Designing an Efficient Reward Function for Robot Reinforcement Learning of The Water Bottle Flipping Task,” *Journal of Korea Robotics Society*, vol. 14, no. 2, pp. 81-86, Jun., 2019, DOI: 10.7746/jkros.2019.14.2.081.

[9] P. Abbeel, M. Quigley, and A. Y. Ng, "Using Inaccurate Models in Reinforcement Learning," *23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, USA, pp. 1-8, 2006, DOI: 10.1145/1143844.1143845.

[10] M. J. Mataric, "Reward Functions for Accelerated Learning," *Eleventh International Conference*, Brunswick, NJ, USA, pp. 181-189, 1994, DOI: 10.1016/B978-1-55860-335-6.50030-1.

[11] H. Hachiya, J. Peters, and M. Sugiyama, "Reward-Weighted Regression with Sample Reuse for Direct Policy Search in Reinforcement Learning," *Neural Computation*, vol. 23, no. 11, pp. 2798-2832, 2011, DOI: 10.1162/NECO_a_00199.

[12] B. C. da Silva, G. Baldassarre, G. Konidaris, and A. Barto, "Learning parameterized motor skills on a humanoid robot," *2014 IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, pp. 5239-5244, 2014, DOI: 10.1109/ICRA.2014.6907629.

[13] S. H. Lee, "Designing an efficient reward function for robot reinforcement learning of the water bottle flipping task," M.S thesis, Sogang University, Seoul, Korea, 2018, [Online], https://library.sogang.ac.kr/search/detail/CAT000000843771.

[14] J. Kober and J. Peters, "Learning Motor Primitives for Robotics," *2009 IEEE International Conference on Robotics and Automation*, Kobe, Japan, pp. 2112-2118, 2009, DOI: 10.1109/ROBOT.2009.5152577.

[15] J. Wang, Y. Liu, and B. Li, "Reinforcement Learning with Perturbed Rewards," *AAAI Technical Track: Machine Learning*, 2020, DOI: 10.1609/aaai.v34i04.6086.

[16] K. Främling, "Reinforcement Learning in a Noisy Environment: Light-Seeking Robot," *WSEAS Transactions on Systems*, vol. 3, no. 2, pp. 714-719, 2004, [Online], https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.484.6001&rep=rep1&type=pdf.

양 영 하

2015 서강대학교 기계공학과 학사
2020 서강대학교 기계공학과 석사

관심분야: Reinforcement Learning

이 철 수

1984 한양대학교 산업공학과 학사
1986 KAIST 산업공학과 석사
1990 KAIST 산업공학과 박사
현재 서강대학교 기계공학과 교수

관심분야: CAD/CAM, CNC controller, 공작기계, 로봇