RPO 기반 강화학습 알고리즘을 이용한 로봇 제어

Robot Control via RPO-based Reinforcement Learning Algorithm

김종호, 강대성, 박주영

고려대학교 제어계측공학과

Jongho Kim, Daesung Kang, Jooyoung Park

Dept. of Control & Instrumentation Engineering, Korea University

E-mail: oyeasw@korea.ac.kr, mpkds@korea.ac.kr, parkj@korea.ac.kr

Abstract

The RPO algorithm is a recently developed tool in the area of reinforcement learning. And it has been shown to be very successful in several application problems. In this paper, we consider a robot-control problem utilizing a modified RPO algorithm, in which its critic network is adapted via RLS(Recursive Least Square) algorithm. We also developed a MATLAB-based animation program, by which the effectiveness of the training algorithms were observed.

키워드 : 강화학습, RPO 알고리즘, Kimura의 로봇

1. 서론

강화 학습은 기계학습(machine learning) 분야의 주요한 도구로써 여러 분야에서 흥미 있는 결과를 계속적으로 제공하여 왔는데, 최근에는 자동제어 관련 분야에서도 흥미 있는 적용 사례가 보고된 바 있다[4]. 강화학습은 지도학습과 비지도 학습의 중간적인 특성을 가지고 있어서 시도와 오류(trial-error)를 통해서 정책이 결정되기 때문에 구체적인 모델이 필요 없는 장점을 가지고 있다. 한편 강화학습에는 가치 반복(value iteration)을 이용하는 학습과 정책 반복(policy iteration)을 이용하는 학습이 있는데, 본 논문에서 다루고자 하는 RPO방법은 후자에 속하는 방법이다.

정책 반복을 이용하는 방법 중 하나에는 actor-critic 방법이 있는데, 이들은 actor와 critic에 대한 학습을 필요로 한다. critic 학습은 정책의 실행에 관련된 부분으로 현재 상태와 다음상태의 가치의 차에 의해서 계산되며[5], 계산된 값들은 actor의 행동결정에 사용된다. actor 학습은 정책의 조정과 관련된 부분으로 최적의 제어 입력을 선택하는 과정이다. 본 논문에서는 Wawrzynski[1],[2] 등에 의해서 소개된 RPO(λ) 알고리즘을 이용하되, critic 부분에 RLS(Recursive Least Square)[3]를 적용한후, Kimura[4]등에 의해서 소개된 기는 로봇에 이 알고리즘을 적용하여 보았다.

본 논문의 구성은 다음과 같다. 2장에서는, 본 논문의 주요 소재가 되는 Kimura의 로봇에 대하여 간단히 설명한 후, 연속 공간에서 SGA(stochastic gradient ascent)를 적용한 예가 소개된다. 3장에서는 본 논문의 주된 관심사인 RPO(λ)에 대한 관련 수식과 제어 입력, 그리고 가치 함수의 결정에 대해서 언급한 후 로봇에 적용했을 경우에 대한 결과를 설명한다. 마지막으로, 4장에서는 결론과 향후 연구 방향 등을 제시한다.

2. Kimura의 로봇과 학습

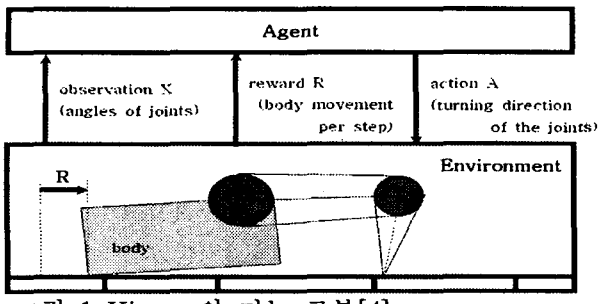

그림 1 Kimura의 기는 로봇[4]

참고문헌 [4]에서 Kimura 등은 강화학습의 효용성을 보이기 위해 간단한 기는 로봇을 응용 문제로 고려하였다. 이 로봇은, 중력이 가해지는 환경 아래에서 두 개의 링크를 가지고 기는 동작을 수행하는 평면형 머니퓰레이터(planar manipulator)로써 그림 1의 구조를 갖는다.

이 로봇에 부과된 임무는 최대한 빨리 전진하는 것인데, 에이전트(agent)는 로봇 및 환경에 대한 구체적인 모델 또는 정보가 주어지지 않은 상태에서 직접적인 경험을 통해 관찰된 보상값(rewards) r 만을 가지고 효과적인 제어 규칙을 발견해내야 한다. 각 시간 스텝 때마다 에이전트는 조인트의 각도를 읽어 들이고 확률적 제어입력 선택 전략에 따라 조인트에 연결된 모터의 회전 방향 및 회전각도를 결정한다.

그리고, 학습 과정에서 이용되는 보상값 r 을 위해서는 해당 시간 스텝 동안 전진한 거리가 사용된다. 만일 로봇이 후진하는 경우에는 후진한 거리만큼의 음의 보상값(negative reward)이 생성됨은 물론이다. 직관적으로 생각할 때에, 위의 로봇이 최대한 빨리 전진하기 위해서는 기면서 앞으로 나아가는 패턴을 신속하게 습득해야 함을 알 수 있다. 본 논문에서 고려하는 로봇 관련 데이터는 [4]의 경우와 같다.

로봇의 위쪽 팔의 길이는 34 cm이고(이하, 단위 생략), 아래쪽 팔의 길이는 20이다. 그리고, 몸체와 위쪽 팔을 잇는 첫 번째 조인트는 몸체의 좌측하단 코너로부터 수평방향으로 32, 수직방향으로 18 떨어진 곳에 위치한다. 몸체와 위쪽 팔을 잇는 조인트의 움직임은 몸체와 수평인 방향에서 [-4, 35] 도 범위에서만 가능하고, 위쪽 팔과 아래쪽 팔을 잇는 두 번째 조인트의 움직임은 위쪽 팔과 수평인 방향에서 [-120, 10] 도 범위에서만 가능하다. 그리고 아래쪽 팔의 뾰족한 끝부분이 지면에 닿아 있을 때에는, 뾰족한 끝부분은 미끄러지지 않고 몸체만 미끄러짐을 가정한다.

그림 2는 [4]의 방법론을 중심으로 매트랩 프로그램을 이용하여 연속시간에서의 SGA[7]를 사용한 결과이다. 학습이 진행됨에 따라 로봇의 평균 진행 속도가 점차적으로 증가하는 패턴을 보여준다.

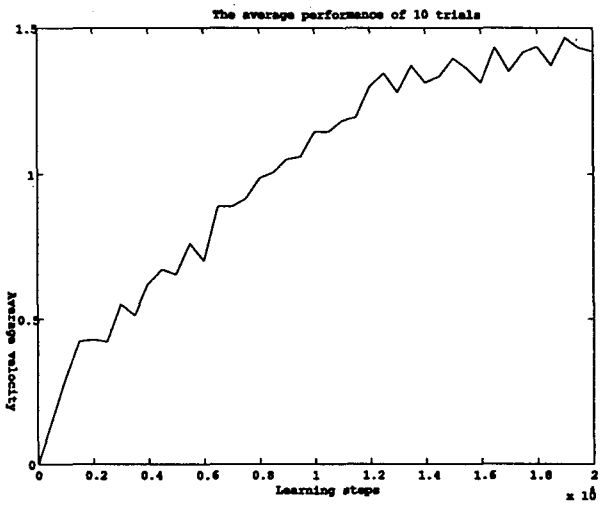

그림 2. 연속제어입력을 갖는 Kimura의 로봇을 SGA 기법으로 학습시킨 결과[7]

3. RPO(λ) 알고리즘을 이용한 학습 및 적용

[1]과[2] 등에서 시도된 RPO(λ) 기법은, Cart-Pole 문제 및 도립 진자 제어 문제 등에 적용된바 있다. 본 논문에서는 [1],[2]에서 언급한 방법론의 RPO(λ)의 critic부분에 RLS를 적용한 결과를 Kimura의 로봇을 대상으로 적용하였다. RPO(λ)-RLS는 다음과 같은 절차로 구성된다.

(1) 파라미터를 초기화함($P_o = \delta I$, P_o= 초기 분산 매트릭스, δ=양의 정수, I=항등 매트릭스)
(2) 시간 스텝 t 때의 관측변수 x_t를 관찰함
(3) 확률분포 $\varphi(.;\tilde{\theta}(x_t;w_\theta))$에 따라, 제어입력 μ_t를 샘플링하여 실행함 (w_θ는 actor의 연결강도)
(4) 행동에 따른 보상값 r_t를 관찰함
(5) 가치함수의 차를 계산함
$,d_t = r_{t+1} + \gamma \tilde{V}(x_{t+1};w_V) - \tilde{V}(x_t;w_V)$
(w_V는 critic의 연결강도)
(6) actor:
a. actor의 적격성(e_θ)과 적격성 트레이스(m_θ)를 계산함

$$e_\theta = \frac{d\tilde{\theta}(x_t;w_\theta)}{dw_\theta}\frac{d\ln\varphi(u_t;\tilde{\theta}(x_t;w_\theta))}{d\tilde{\theta}(x_t;w_\theta)}$$

$m_\theta = \gamma\lambda m_\theta + e_\theta$ (여기에서, $\gamma\in[0,1)$은 할인율을 $\lambda\in[0,1)$ 감쇠율을 나타냄)

b. actor의 연결강도를 개선함

$w_\theta = w_\theta + \beta_t^\theta d_t m_\theta$

(6) critic:

a. critic의 기저함수 벡터(ϕ)과 적격성 트레이스(z)를 계산함

$z = \gamma^*\lambda^* z + \phi$

b. critic의 연결강도 w_V를 개선함

$$K_{t+1} = P_t z_t/(\mu + (\phi^T(x_t)) - \gamma\phi^T(x_{t+1}))P_t z_t)$$

$$W_{v_{t+1}} = W_{V_t} + K_{t+1}(r_t - (\phi^T(x_t) - \gamma\phi^t(x_{t+1}))W_{V_t}$$

$$P_{t+1} = \frac{1}{\mu}(P_t - P_t z_t(1 + (\phi^T(x_t) - \gamma\phi^T(x_{t+1})P_t z_t))^{-1}\times$$
$$\times(\phi^T(x_t) - \gamma\phi^T(x_{t+1})P_t)$$

(7) 시간스텝을 $t+1$로 증가시키고, 단계 (1)로 되돌아감

본 논문에서는 [1]에서의 이론 전개를 참고하여, σ에는 1의 값을 actor에는 각 조인트의 제어입력 선택 전략을 위한 확률분포 φ로 다음과 같은 정규분포를 고려하였다:

$$\varphi(\mu;\theta) = \frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(\mu-\theta)^2}{2\sigma^2})$$

따라서, 평균 μ에 대한 적격성은 다음과 같다.

$$m_\theta = \frac{d\tilde{\theta}(x_t;w_\theta)}{dw_\theta}\sigma^{-2}(\mu_t - \tilde{\theta}(x_t;w_\theta))$$

그리고, 각 조인트에 대한 확률적 제어입력선택 전략 π의 평균 μ를, $\mu = w_{\theta 1}\theta_1 + w_{\theta 2}\theta_2 + w_{\theta 3}$ 의 값으로 선택하면 연결강도의 적격성을 다음과 같이 구할 수 있다.

$$m_{\theta 1} = \theta_1(\mu_t - \tilde{\theta}(x_t;w_\theta))\sigma^{-2}$$
$$m_{\theta 2} = \theta_2(\mu_t - \tilde{\theta}(x_t;w_\theta))\sigma^{-2}$$
$$m_{\theta 3} = (\mu_t - \tilde{\theta}(x_t;w_\theta))\sigma^{-2}$$

두 번째 조인트를 위한 제어입력 연결강도의 적격성 역시 비슷한 방법으로 구해진다. 그리고 각 조인트에서는 로봇의 과도한 움직임을 막기위해서 각 시간 스텝 당 [-12도 , 12도]범위까지의 움직임만 허용하는 한계성을 부여하였다. 위의 식들에 등장하는 θ_1과 θ_2는, 각 조인트의 각도 변위가 [-1,1] 범위가 되도록, 관련 측 변수인 조인트를 적절하게 스케일링한 결과로 정의되는 관측변수이다.

일반적으로 RLS는 학습 속도가 빠른 것이 장점이다. critic 부분에서는 RLS를 이용하여 하중벡터를 학습시키고, 학습된 하중벡터는 가치함수를 근사화 한다. 근사화된 값은 actor의 행동결정 방법에 이용된다.

일반적인 기저함수는 아래와 같이 표현된다.

$\phi(x) = (\phi_1(x),\phi_2(x),\phi_3(x),\phi_4(x),\phi_5,\phi_6(x))^T$

본 논문에서는 $\phi(x)$를 관측변수 θ와 동일하게 사용하였다. 기저함수를 이용한 가치근사($\tilde{V}$)는 다음과 같다.

$\tilde{V_t}(x) = \phi^T(x)W_{V_t}$, $\tilde{V_{t+1}}(x) = \phi^T(x_{t+1})w_{V_{t+1}}$

$W_V = (w_1,w_2,w_3,w_4,w_5,w_6)^T$

이때의 적격성 트레이스는 $z = \gamma\lambda z + \phi(x_t)$ 를 사용하여 각각 조인트에 대한 적격성을 고려하였다.

학습에서 사용된 그 밖의 관련 파라메터는 다음과 같다. 할인율 $\gamma = 0.9$, 감쇠율 $\lambda = 0.75$, 학습율 $\beta_t^\theta = 0.006$, 초기 분산상수 $\delta = 0.2$

이상에서 설명한 제어입력을 갖는 Kimura의 로봇에 RPO(λ)-RLS 학습기법을 적용하여 매트랩 시뮬레이션을 수행한 결과는 그림 3과 같고 2장의 경우보다 우수한 성능이 관찰되었다. 또한 개발된 시뮬레이터를 바탕으로, 향후 학습효과를 시각적으로 관찰하는데 큰 도움이 될 것이다.

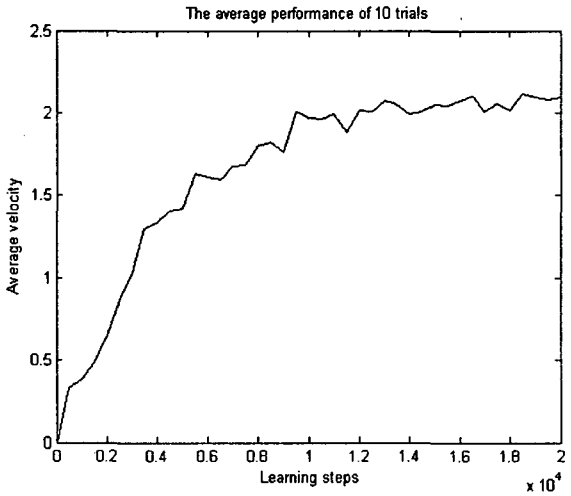

그림 3. Kimura의 로봇을 RPO(λ)-RLS를 적용하여 학습시킨 결과

4. 결론 및 향후과제

본 논문에서는 Kimura의 로봇을 대상으로 하여,

RPO(λ)의 critic 부분에 RLS를 적용한 문제를 고려해보았다. 로봇의 제어는 [4]에서 고려된 이산제어 입력, 연속제어 입력을 위한 SGA 기법보다 RPO(λ)에 RLS를 접목시킨 것이 우수한 효과를 보임을 관찰하였다. 강화학습 분야에 여러 가지 흥미 있는 새로운 알고리즘이 꾸준히 제안되고 있는 현실을 생각할때, 본 연구를 통해 확보된 시뮬레이터를 바탕으로 여러 강화학습 알고리즘의 효과를 비교, 관찰해 볼 수 있는 좋은 도구가 될 것이라 생각한다. 향후에 시도해 볼만한 연구로는, 최근 기계학습 분야에 큰 영향을 미치고 있는 커널 기법을 강화 학습 분야에 접목시킨 학습 알고리즘 개발 후 이를 시뮬레이터를 통해 확인해 보는 문제 등을 들 수 있다.

참고문헌

[1] P. Wawrzynski and A. Pacut, "A simple actor-critic algorithm for continuous environments," *Proceedings of the 10th IEEE Int. Conf. on Methods and Models in Automation and Robotics, Miedzyzdroje, Poland,* August 2004, pp. 1143-1149.

[2] P. Wawrzynski and A. Pacut, "Model-free off-policy reinforcement learning in continuous environment," *Proceedings of the International Joint Conference on Neural Networks, Budapest, July 2004,* pp. 1091-1096.

[3] X. Xu, H. He and D. Hu, "Efficient Reinforcement Learning Using Recursive Least-Square Methods," Journal of Artificial Intelligence Research, vol 16, pp. 259-292, 2002

[4] H. Kimura, K. Miyazaki, and S. Kobayashi, "Reinforcement learning in POMDPs with function approximation," In *Proceedings of the 14th International Conference on Machine Learning (ICML '97),* pp. 152-160, 1997.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction,* MIT Press, 1998.

[6] H. Kimura and S. Kobayashi, "An Analysis of Actor/Critic Algorithms using Eligibility Traces: Reinforcement Learning with Imperfect Value Function," *15th International Conference on Machine Learning,* pp.278--286 (1998).

[7] 박주영, 김종호, 신호근, "SGA 기반 강화학습 알고리즘을 이용한 로봇 제어" 한국 퍼지 및 지능시스템 학회 2004년도 추계학술 대회 논문집, 14권 2호, pp. 63-66, 2004년 10월