2005년도 대한전기학회 하계학술대회 논문집 2005. 7. 18 - 20

IRPO 기반 Actor-Critic 학습 기법을 이용한 로봇이동

김종호, 강대성, 박주영
고려대학교 제어계측공학과

Robot locomotion via IRPO based Actor-Critic Learning Method

Jongho Kim, Daesung Kang, Jooyoung Park
Dept. of Control & Instrumentation Engineering, Korea University

Abstract - The IRPO(Intensive Randomized Policy Optimizer) 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. To compare with a general RL method, IRPO has some difference in that policy utilizes the entire history of agent-environment interaction. The policy is derived from the history directly, not through any kind of a model of the environment. In this paper, we consider a robot-control problem utilizing a IRPO algorithm. We also developed a MATLAB-based animation program, by which the effectiveness of the training algorithms were observed.

1. 서 론

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[5]. 본 논문에서 다루고자 하는 IRPO는 일반적인 학습방법과 달리 다루고자 하는 대상(environment)이 경험한 모든 상태정보를 이용하여 제어 입력을 선택하는 학습기법이다. 한편 강화학습에는 가치 반복(value iteration)을 이용하는 학습과 정책 반복(policy iteration)을 이용하는 학습이 있는데, 본 논문에서 다루고자 하는 IRPO방법은 후자에 속하는 방법이다.

정책 반복을 이용하는 방법은 정책의 실행과 개선으로 이루어져 있으며, actor-critic 방법이 이에 속한다. 이들은 actor와 critic에 대한 학습을 필요로 한다. 한편 본 논문의 주된 관심인 IRPO는 이전 스텝의 모든 data를 이용하여 actor와 critic을 학습시킨다.

critic의 학습은 정책 실행과정에서 나타나는데, 가치 함수 근사를 통해 나타나는 에러를 최소화하기 위한 하중 벡터의 개선을 통해 학습이 이루어진다. 한편 actor의 학습은 정책 개선 과정에 나타나며, 최적의 제어 입력을 선택하기 위한 하중 벡터 개선과정으로 표현된다.

본 논문에서는 Wawrzynski[1],[2] 등에 의해서 소개된 IRPO 알고리즘을 이용하여 Kimura[4]의 로봇을 대상으로 하여, IRPO알고리즘의 성질을 확인해 보는 투토리얼 논문의 성질을 가지고 있다.

본 논문의 구성은 다음과 같다. 2.1장에서는, 본 논문의 주요 소재가 되는 Kimura의 로봇에 대하여 간단히 설명한 후, 연속 공간에서 RPO(λ)-RLS 적용한 예가 소개된다. 2.2장에서는 본 논문의 주된 관심사인 IRPO에 대한 관련 수식과 제어입력, 그리고 정책 반복과 개선에 대해서 언급한 후 로봇에 적용했을 경우에 대한 결과를 설명한다. 마지막으로 결론과 향후 연구 방향 등을 제시한다.

2. 본 론

2.1 Kimura의 로봇과 학습

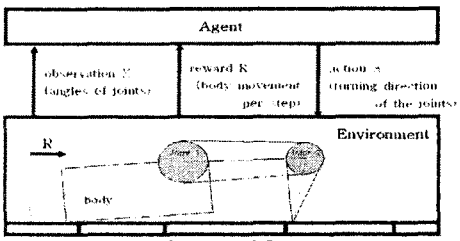

그림 1 Kimura의 기는 로봇[4]

참고문헌 [4][6]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 머니플레이터(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 회전 각도를 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.

2.2 IRPO 알고리즘을 이용한 학습 및 적용

그림2는 [2]의 방법론을 존중하면서, critic 부분에

RLS(Recursive Least Square)을 적용해 학습시킨 결과를 나타내고 있다. 시도와 에러를 통해서 학습이 진행됨에 따라 로봇의 평균 진행 속도가 점차적으로 증가하는 패턴을 보여준다.

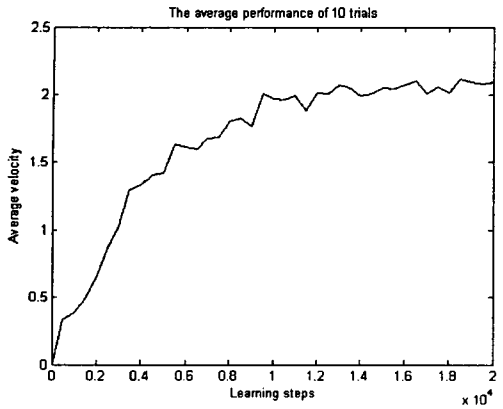

그림 2. Kimura의 로봇에 RPO(λ)-RLS를 적용하여 학습시킨 결과[7]

[1]에서 시도된 IRPO 기법은, Cart-Pole 문제 및 도립진자 제어 문제 등에 적용된바 있다. 본 논문에서는 IRPO 방법론을 Kimura의 로봇을 대상으로 적용하였다. IRPO 알고리즘은 $\mathcal{B}$와 V의 2개의 파라미터 근사로 구성되어 있는데, $\mathcal{B}$는 확률적 제어 정책(randomized control policy) $\pi(w_\theta)$를 근사화하는데 사용되며 , V의 값은 $V^{\pi(\theta)}$의 값을 근사화 하는데 사용된다.

한편 IRPO 학습 알고리즘은 다음의 2개 loop로 구성되어 있다.

a. Exploring loop : agent는 environment와 상호 작용을 통해서 history를 수집한다.
b. Internal loop : agent는 그 동안의 history를 바탕으로 정책을 최적화 한다.

IRPO의 exploration loop는 다음의 step으로 구성된다.

1) $\mu_t \sim \varphi(.,\mathcal{B}(s_t;w_\theta))$를 통한 제어 입력 μ_t의 선택, (w_θ는 internal loop에서 계산된 하중벡터)
2) 제어 입력의 실행, 보상값(r_{t+1})과 다음 상태 (s_{t+1}) 관찰
3) $\langle s_t, u_t, r_{t+1}, s_{t+1}, \mathcal{B}(s_t;w_\theta)\rangle$의 값을 dataset에 저장
4) step 1로 이동 후 반복

한편 internal loop는 다음의 step으로 구성된다.

1) 정책 실행: 가치 함수 $V(s_i;w_V)$ 결정하기위한 하중벡터 w_V의 개선
2) 정책 개선: 목적 함수 $U^\pi(s_i,\mathcal{B}(s_i;w_\theta))$의 값을 최대화 하기 위한 하중벡터 w_θ의 개선

internal loop의 2개의 step은 정책 반복을 실행하는 과정이다. 첫번째 step 은 현재 정책의 가치 함수(Value Function)를 계산하고, 두번째 step은 각 history에서 첫번째 step을 최적화 하는 과정이다. 위의 내용을 바탕으로 하는 IRPO 학습 알고리즘은 다음과 같다.

1) 확률 분포 $\varphi(.,\mathcal{B}(s_t;w_\theta))$에 따라, 제어 입력 μ_t를 선택
2) 다음 상태 s_{t+1} 와 보상값 r_{t+1} 관찰
3) $\langle s_t, u_t, r_{t+1}, s_{t+1}, \mathcal{B}(s_t;w_\theta)\rangle$를 각각 저장
4) (The internal loop) repeat :
   a. $d_i = r_{t_i} + \gamma V(s_{t_{i+1}};w_V) - V(s_{t_i};w_V)$ 계산
   b. $w_V := w_V + \beta_i^V d_i \rho_b(u_{t_i}, \mathcal{B}(s_{t_i};w_\theta), \theta_{t_i}) \times \frac{dV(s_{t_i};w_V)}{dw_V}$, β_i^V: critic학습율
   c. 정책 개선 :

$$g_i = G\left(\frac{d\rho_b(u_{t_i}, \mathcal{B}(s_{t_i};w_\theta), \theta_{t_i})}{d\mathcal{B}(s_{t_i};w_\theta)}, \mathcal{B}(s_{t_i};w_\theta)\right)$$

$$w_\theta := w_\theta + \beta_i^\theta d_i \frac{d\mathcal{B}(s_{t_i}, w_\theta)}{dw_\theta} g_i \ , \quad \beta_i^\theta = actor\text{학습율}$$

   d. Set $i := i+1$
5) $t = t+1$, step 1을 반복

본 논문에서는 [1][3]에서의 이론 전개를 참고를 하여 C=4의 값을로 설정하였다. 그리고 각 조인트의 제어 입력 선택 전략을 위한 확률 분포 φ로 다음과 같은 정규 분포를 고려하였다.

$$\varphi(u,\theta) = \frac{1}{\sqrt{2\pi C}} \exp\left(-\frac{1}{2}(\mu-\theta)^T C^{-1}(\mu-\theta)\right)$$

IRPO의 확률적 제어입력 선택은 $\theta: X \rightarrow U$를 통한 제어 θ를 최적화하기 위한 과정이다. 이는 현재 상태의 값에서 결정되는 평균값을 제어입력으로 바로 이용하는 것이 아니라, 분산에 의한 noise를 포함하는 형식으로 표현된다. 이러한 noise의 값은 최적제어 탐색 과정에서, 다양한 action을 취하도록 함으로써 최적의 해를 찾는 과정에 속한다. 최적 제어입력 과정에서 이러한 noise는 반드시 필요하며, variance의 값을 통해서 제어입력의 탐색과 이용의 정도를 조절 할 수 있다.

각 조인트에 대한 확률적 제어 입력 선택 전략 π의 평균 μ는, $\mu = w_{\theta 1}\theta_1 + w_{\theta 2}\theta_2 + w_{\theta 3}$의 값을 선택했으며, 두 번째 조인트를 위한 제어 입력의 선택 역시 비슷한 방법으로 구해진다. 그리고 각 조인트에서는 로봇의 과도한 움직임을 막기 위해[-12도, 12도]범위까지의 움직임을 허용하는 한계성을 부여하였다. 위의 식에 등장하는 θ_1과 θ_2는 , 각 조인트의 각도 변위를 [-1,1]범위가 되도록 관련 측 변수인 조인트 각도를 적절하게 스케일링한 결과로 정의되는 관측 변수 이다.

t_i의 값은 history에서 random하게 선택된 dataset의 index를 나타낸다.

한편 충분한 dataset을 확보하기 위해 스텝이 작은 경우(스텝<100)에는 internal loop 없이 external loop 만을 반복했으며, 스텝이 큰 경우(스텝>100)인 경우부터 internal loop를 시작했다. internal loop의 반복 횟수를 나타내는 n의 값은 처음에 $n=10$ 값을 시작으로 internal loop를 반복하면서 점차 증가하여, 1000을 넘지 않도록 했다. internal loop에서 하중 벡터 w_θ, w_V를 개선하기 위한 과정에 나타나는 ρ_b의 값은 학습 과정의 variance를 줄이기 위해서 사용되는 estimator로 다음과 같이 표현 가능하다.

$$\rho_b(Y_0, \theta, \theta_0) = \min\left(\frac{\varphi(Y_0,\theta)}{\varphi(Y_0,\theta_0)}, b\right),\ b=5$$

한편, $G(h,\theta)$의 값은 $\mathcal{H}(s_t;w_\theta)$를 일정한 범위 안에 구속시키기 위해서 사용되는 값으로 다음과 같이 나타낼 수 있다.

$$G(h,\theta)=h$$

그 밖에 실험에서 사용된 파라미터의 값은 다음과 같다.
β_i^θ=0.002 , β_i^V=0.001

2.1장과 2.2장에서 언급한 내용을 바탕으로 Kimura의 기는 로봇을 대상으로 하는 실험 결과는 다음과 같다. 로봇의 2개 joint에 대한 회전방향과 각도는 actor의 파라미터(w_θ)와 2개 조인트에 대한 각도를 [-1,1]로 스케일링 벡터의 선형 결합($\theta=\sum_i w_{\theta i}{}^T\phi_i$)으로 결정된다. 한편 제어입력이 매 step 마다 결정되는(deterministic) 한 경우가 아닌 확률적 선택(stochastic)한 경우이기에, 각 조인트에 대한 최종 제어입력은 $\mu_t(.;)\sim N(\theta,\sigma^2)$으로 구현된다.

모두 10번의 episode를 실행했으며, 각 episode는 20000번의 step으로 구성되어 있다. 평균속도는 500step의 배수에 그동안 학습된 actor의 파라미터를 이용하여, 이동한 거리를 500으로 나눈값으로 했으며, 모든 episode의 평균속도를 합하여 그에 대한 평균을 각 step의 최종평균으로 하였다.

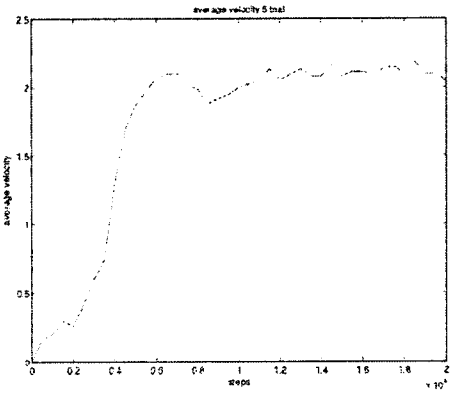

그림 3. Kimura의 로봇에 IRPO를 적용하여 학습시킨 결과

3. 결 론

본 논문에서는 Kimura의 로봇을 대상으로 하여, IRPO를 이용하여 학습 시켰을 경우에 대한 성능을 RPO (λ)-RLS와 비교하였다. IRPO을 이용한 학습방법이 RPO(λ)-RLS의 방법보다 우수한 효과를 보임을 관찰하였다. IRPO는 기존의 학습 방법과 달리 과거에 가지고 있던 모든 상태의 정보를 이용하여 확률적으로 기존 dataset을 샘플링하여 학습에 이용하는 방법이다. 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. 그리고 최근 critic을 위한 특정한 함수 근사와 기울기 강하(gradient descent) 방법을 결합하여 현재의 IRPO의 성능보다 개선된 학습 알고리즘을 개발하는 문제 역시 고려하고 있다.

[참 고 문 헌]

[1] 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.

[2] P. Wawrzynski and A. Pacut, "Intensive versus non-intensive actor-critic reinforcement learning algorithms", Proceedings of the 7th International Conf. on Artifical Intelligence and Soft Computing, Poland, June 2004, pp. 934-941,

[3] D. Precup, R. S. Sutton, S. Singh, *"Eligibility Traces for Off-Policy Policy Evaluation," Proceedings of the 17th international Conference on Machine Learning*, Morgan Kaufman, 2000

[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] 김종호,강대성,박주영 "RPO 기반 강화학습 알고리즘을 이용한 로봇 제어" 한국 퍼지 및 지능시스템 학회 2005년도 춘계학술 대회 논문집 15권, 1호, 217-220