퍼지 및 지능시스템학회 논문지 2005, Vol. 15, No. 4, pp. 505-510

RPO 기반 강화학습 알고리즘을 이용한 로봇제어

Robot Control via RPO-based Reinforcement Learning Algorithm

김종호, 강대성, 박주영
Jongho Kim, Daesung Kang and Jooyoung Park

고려대학교 제어계측공학과

요 약

제어 입력 선택 문제에 있어서 확률적 전략을 활용하는 RPO(randomized policy optimizer) 기법은 최근에 개발된 강화학습 기법으로써, 많은 적용 사례를 통해서 그 가능성이 입증되고 있다. 본 논문에서는, 수정된 RPO 알고리즘을 제안하는데, 이 수정된 알고리즘의 크리틱 네트워크 부분은 RLS(recursive least square) 기법을 통하여 갱신된다. 수정된 RPO 기법의 효율성을 확인하기 위해 Kimura에 의해서 연구된 로봇에 적용하여 매우 우수한 성능을 관찰하였다. 또한, 매트랩 애니메이션 프로그램의 개발을 통해서, 로봇의 이동이 시간에 따라 가속되는 학습 알고리즘의 효과를 시각적으로 확인 할 수 있었다.

Abstract

The RPO(randomized policy optimizer) algorithm, which utilizes probabilistic policy for the action selection, is a recently developed tool in the area of reinforcement learning, and has been shown to be very successful in several application problems. In this paper, we propose a modified RPO algorithm, whose critic network is adapted via RLS(Recursive Least Square) algorithm. In order to illustrate the applicability of the modified RPO method, we applied the modified algorithm to Kimura's robot and observed very good performance. We also developed a MATLAB-based animation program, by which the effectiveness of the training algorithms on the acceleration of the robot movement were observed.

Key words : 강화학습(Reinforcement Learning), RPO 알고리즘, Actor-Critic, Kimura의 로봇

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[4][6]. 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.

정책반복을 이용하는 방법 중 하나에는 actor-critic 방법이 있는데, 이들은 actor와 critic에 대한 학습을 필요로 한다. Critic 학습은 정책의 실행에 관련된 부분으로 일반적으로 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]등 에 의해서 소개된 기는 로봇에 이 알고리즘을 적용하여 보았다.

본 논문의 구성은 다음과 같다. 2장에서는 강화학습의 학습방법중 하나인 actor-critic 방법에 대해서 소개를 하며, 3장에서는, 본 논문의 주요 소재가 되는 Kimura의 로봇에 대하여 간단히 설명한 후, 연속 공간에서 SGA(stochastic gradient ascent)를 적용한 예가 소개된다. 4장에서는 본 논문의 주된 관심사인 RPO(λ)-RLS에 대한 관련 수식과 제어입력, 그리고 가치 함수의 결정에 대해서 언급한다. 수정된 RPO 기법을 로봇에 적용했을 경우에 대한 결과를 5장에서 설명하고 마지막으로, 6장에서는 결론과 향후 연구 방향 등을 제시한다.

접수일자 : 2005년 4월 1일
완료일자 : 2005년 6월 30일

2. Actor-critic 알고리즘

일반적으로 강화학습은 제어의 대상이 되는 환경(environment)이 가지고 있는 현재의 상태(x_t)와 상태에 따른 제어입력(a_t) 그리고 이러한 제어입력 후에 관찰되는 보상값(r_t)로 구성되어 있다. 강화학습은 보상값을 최대로 얻기 위한 방향으로 파라미터 개선을 통해서 학습이 이루어 지는데, actor-critic 방법은 제어입력을 선택하는 actor 부분과 제어입력에 대한 적절성 여부를 관찰하는 critic으로 구성되어 있으며, 이들 모두는 각각의 파라미터 벡터를 가지고 있다.

2.1 Actor의 학습

일반적으로 가치함수(Value function)는 현재상태에서 제어입력을 취했을 경우 향후에 거두게 되는 보상값의 기대값으로 정의된다.

$$V^{\pi}(x) = E(\sum_{i \geq 0} \gamma^{i} r_{t+1} | x_t = x; \pi) \tag{2.1}$$

한편 (2.1)의 가치함수는 확률밀도 $d\eta$를 갖는 형태로 표현될 수 있다.

$$\Phi(w_\theta) = \int V^{\pi(w_\theta)}(x)\, d\eta(x, w_\theta) \tag{2.2}$$

한편 입력-가치 함수(action-value function)은 아래와 같이 정의된다.

$$\begin{aligned} U^{\pi}(x, \theta) &= E(r_{t+1} + \gamma V^{\pi}(x_{t+1}) | x_t = x; \mu_t \sim \varphi(.; \theta)) \\ &= E_\theta Q^{\pi}(x, Y) \end{aligned} \tag{2.3}$$

actor는 각 step마다 $\Phi(w_\theta)$를 최대화 하는 방향으로 제어 입력 $\mu_t \sim \varphi(.; \mathcal{V}(x_t; w_\theta))$를 선택하게 되며 그때의 근사함수는 아래와 같다.

$$\nabla_{w_\theta} \Phi(w_\theta) = \frac{d}{dw_\theta} U^{\pi}(x_t, \mathcal{V}(x_t; w_\theta)) \tag{2.4}$$

즉, 가치함수를 최대화하기 위해 $\Phi(w_\theta)$의 기울기(gradient) 방향으로 actor의 파라미터 벡터인 w_θ를 개선하는 방향으로 actor의 학습이 이루어 진다.

2.2 Critic 학습

(2.1)에서 정의한 것처럼 $V^{\pi(w_\theta)}$의 값은 모든 스텝에 대한 기대값으로 표현되며, 이를 critic은 매 step마다 $\mathcal{V}$로 근사하게 된다. 이런 $\mathcal{V}$의 값은 파라미터 벡터 w_V를 포함하고 있으며, critic의 학습은 추정하고자 하는 V와 이에 대한 근사함수 $\mathcal{V}$의 차를 최소화 하는 방향으로 학습이 이루어진다.

$$\Psi(w_V, w_\theta) = \int (V^{\pi(w_\theta)} - \mathcal{V}(x, w_V))^2 d\eta(x, w_\theta) \tag{2.5}$$

RPO를 이용한 입력-가치 함수는 확률적 제어입력에 선택에 의존하므로 확률밀도 함수 $d\eta$를 도입하여 아래와 같이 표현가능하다.

$$\Phi(w_\theta) = U^{\pi(w_\theta)}(x, \mathcal{V}(x; w_\theta)) d\eta(x, w_\theta) \tag{2.6}$$

즉, critic은 아래의 수식에 따른 가치 함수의 차를 최소화하는 방향의 파라미터를 개선하는 학습이 이루어 진다.

$$\begin{aligned} &\Psi(w_V, w_\theta) \\ &= \int (U^{\pi(w_\theta)}(x, \mathcal{V}(x; w_\theta)) - \mathcal{V}(x; w_V))^2 d\eta(x, w_\theta) \end{aligned} \tag{2.7}$$

본 논문 및 실험에서는 $\mathcal{V}$의 값을 각 step이 갖는 특성벡터(feature vector) $\phi(x_t)$와 w_V의 값을 이용하여 근사화 하였다.

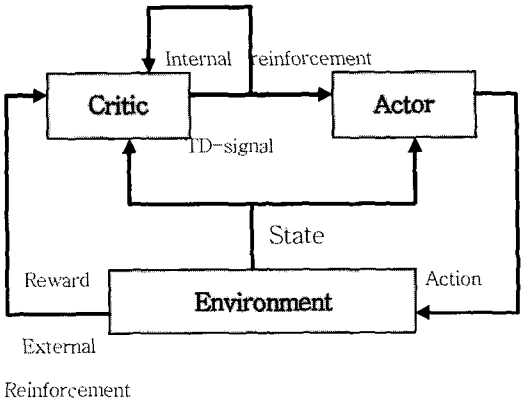

그림 1. Actor-Critic 알고리즘의 구성

그림1은 일반적인 actor-critic 알고리즘이 구성되어 있는 모습을 나타내고 있다. Critic 부분은 TD(temporary difference)방법으로 actor의 제어입력을 평가한다. TD-signal 다음과 같이 정의 된다.

$$r_{t+1} + \gamma \mathcal{V}(x_{t+1}) - \mathcal{V}(x_t) \tag{2.8}$$

이는 제어입력 후에 획득하는 순간 보상값 r_{t+1}와 이후에 거두게 되는 보상값 $V(x_{t+1})$ 이 현재의 보상값 $V(x_t)$보다 큰 경우에 제어입력이 선택될 확률을 증가시키는 방향으로 발생확률을 변화시키게 된다. 반대의 경우에는 제어입력이 선택될 확률을 감소시키는 방향으로 파라미터 개선을 하게 된다. 이러한 actor-critic 학습은 제어입력의 선택에 있어서 최소한의 계산만을 필요로 하며, 확률적 정책을 정확하게 정의할 수 있는 장점을 가지고 있다.

3. Kimura의 로봇과 학습

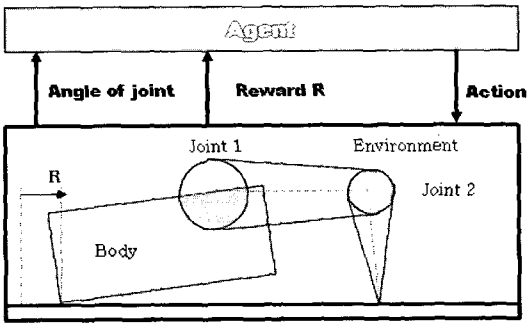

그림 2. Kimura의 기는 로봇[4]

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 관찰

된 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 r을 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 3는 [4]의 방법론을 중심으로 매트랩 프로그램을 이용하여 연속시간에서의 SGA[8]를 사용한 결과이다. 학습이 진행됨에 따라 로봇의 평균 진행 속도가 점차적으로 증가하는 패턴을 보여준다. 평균속도는 매 500step에 그 동안 학습된 actor의 파라미터를 이용해 이동한 거리를 500으로 나눈 값으로 하였다.

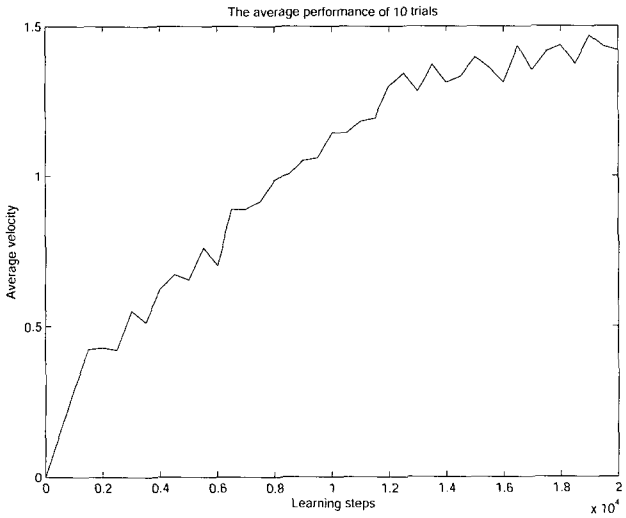

그림 3. 연속제어입력을 갖는 Kimura의 로봇을 SGA 기법으로 학습시킨 결과[8]

4. 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.

2.1절과 2.2절에서 언급한 것처럼 actor는 향후에 거두게 되는 보상값 $\Phi(w_\theta)$의 값을 최대화 하기 위한 방향으로 critic은 $\Psi(w_V, w_\theta)$를 최소화 하는 방향으로 각각 파라미터를 개선시킨다.

Actor와 critic의 학습을 이용한 RPO(λ)-RLS의 학습순서는 다음과 같다.

(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 $\varphi(.;\mathcal{T}(x_t;w_\theta))$에 따라, 제어입력 μ_t를 샘플링하여 실행함 (w_θ는 actor의 연결강도)

(4) 행동에 따른 보상값 r_t를 관찰함

(5) 가치함수의 차를 계산함

$d_t=r_{t+1}+\gamma\mathcal{V}(x_{t+1};w_V)-\mathcal{V}(x_t;w_V)$

(w_V는 critic의 연결강도)

(6) actor:

a. actor의 적격성(e_θ)과 적격성 트레이스(m_θ)를 계산함

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

$m_\theta=\gamma\lambda m_\theta+e_\theta$ (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

$$z_t=\gamma\lambda z_t+\phi$$

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, 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.

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

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 연결

강도의 적격성을 다음과 같이 구할 수 있다.

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

두 번째 조인트를 위한 제어입력 연결강도의 적격성 역시 비슷한 방법으로 구해진다. 그리고 각 조인트에서는 로봇의 과도한 움직임을 막기 위해서 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.

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

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

$$\mathcal{V}_t(x) = \phi^T(x) W_{V_t}, \quad \mathcal{V}_{t+1}(x) = \phi^T(x_{t+1}) w_{V_{t+1}}$$
$$W_V = (w_1, w_2, w_3, w_4, w_5, w_6)^T$$

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

5. 모의 실험

3장과 4장에서 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) 경우가 아닌 확률적 선택(stochastic)한 경우이기에, 각 조인트에 대한 최종 제어입력은 $\mu_t(.;) \sim N(\theta, \sigma^2)$으로 구현된다.

한편 실험에 사용한 다른 파라미터는 다음과 같다.

$\sigma^2 = 1$, $\gamma = 0.9$, 감쇠율 $\lambda = 0.75$, 학습율 $\beta_t^\theta = 0.006$

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

그림 4의 결과는 앞에서 제안한 RPO(λ)-RLS를 이용하여 기는 로봇에 적용한 결과를 나타낸다. 제안한 RPO(λ)-RLS의 방법은 기존의 RPO(λ)의 방법의 critic 학습부분을 RLS로 대체한 방법이다. 제안한 방법이 기존의 방법보다 우수한지를 확인하기 위해서 [1]에서 언급한 RPO 방법과의 성능비교를 하였다.

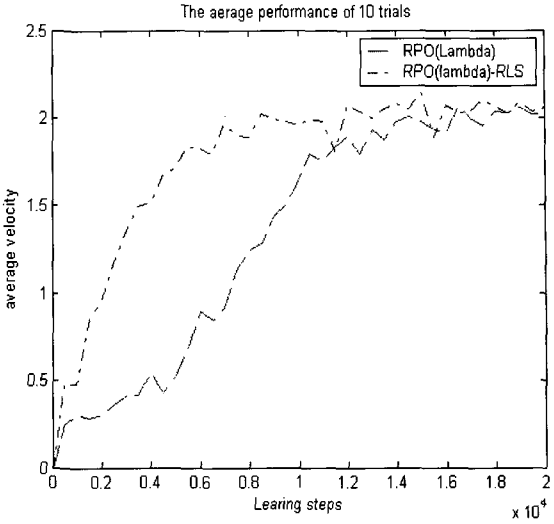

그림 4. 기는로봇에 RPO(λ)-RLS와 RPO(λ)의 학습성능 비교

그림 4에서 볼 수 있는 것처럼 RPO-RLS는 RPO 방법보다 초기 학습 능력이 뛰어나며, 빠른 수렴성을 가지고 있음을 확인 할 수 있다. 물론 그림3의 SGA방법보다 우수함도 확인 할 수 있다.

일반적으로 강화학습은 사용하는 파라미터에 의해서 많은 영향을 받으며 서로 다른 성능을 보인다. 학습율(β_θ^t), 감쇠율(γ)등과 같은 변수의 값을 달리하면 각각 서로 다른 결과를 보이는 것이 일반적인 특징이다.

인용논문 [1]에서 언급 한 것처럼 RLS를 이용한 경우는 critic의 P_0(초기분산 메트릭스)의 값에 의해서 학습 정도가 크게 영향을 받느다. 한편 joint의 최종 제어입력 $N(\theta, \sigma^2)$의 경우에도 σ^2의 값에 의해서 서로 다른 학습 정도를 보임을 아래의 결과에서 확인 할 수 있다.

변수가 미치는 영향을 조사하기 위해서 기존의 값은 그대로 유지하면서 관측하고자 하는 변수들만을 달리하여 최종학습 결과를 관찰하였다.

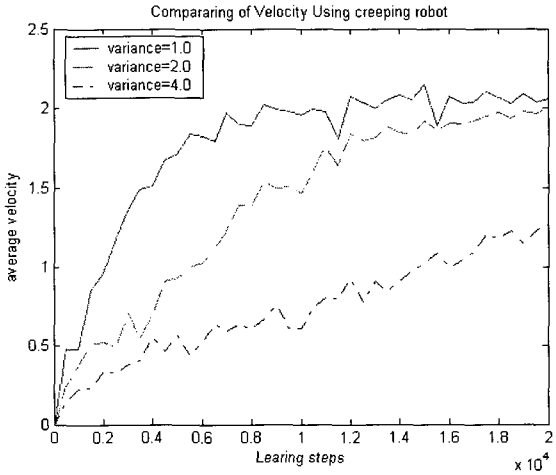

그림 5. RPO(λ)-RLS를 적용, $P_0 = 0.2$,서로 다른 분산을 적용 했을때의 평균속도

RPO의 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.

그래서 variance의 값을 달리 하면 서로 다른 평균속도를 보인다. 최적 제어입력 과정에서 이러한 noise는 반드시 필요하며, variance의 값을 통해서 제어입력의 탐색과 이용의 정도를 조절 할 수 있다. 고정된 분산값을 이용하는데 있어서 큰 분산값을 이용할 경우 빠른 수렴성을 보장하지만 최종적인 수행결과가 취약할 수 있으며, 반대로 작은 분산값을 이용하면 늦은 수렴성을 보장하지만 최종 결과가 우수하다. 즉 적절한 variance를 선택하는 것이 중요하다.

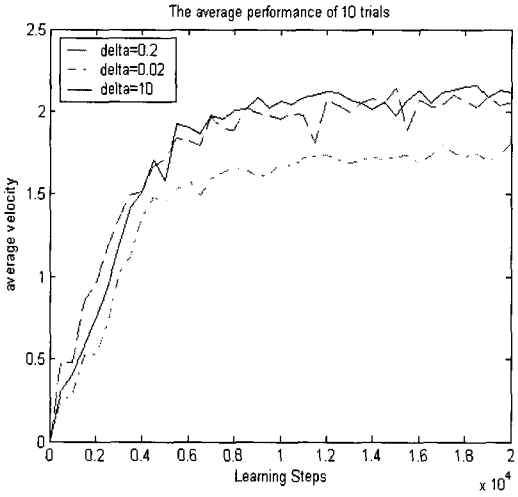

그림 6. RPO(λ)-RLS를 적용 , $\sigma^2=1$, 서로 다른 δ의 값

기는 로봇의 학습을 결정하는 것은 앞에서 언급한 variance 뿐만 아니라 초기 분산 매트릭스의 값에 의해서도 결정이 된다. 위의 그림은 $\delta>0$으로 정의되는 분산 매트릭스의 값을 달리했을 경우에 대한 기는 로봇의 평균속도를 나타낸다.

6. 결론 및 향후 과제

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. 또한 제안한 학습기법인 RPO(λ)-RLS가 기존의 방법인 RPO(λ)보다 효율적인 학습기법임을 확인 할 수 있었다. 모의 실험에서 알 수 있었던 것처럼 학습에 사용하는 계수(coefficient)의 값을 달리 하면 학습 효과에 큰 영향을 미치는 것을 확인 할 수 있었다.

강화학습 분야에 여러 가지 흥미 있는 새로운 알고리즘이 꾸준히 제안되고 있는 현실을 생각할때, 본 연구를 통해 확보된 시뮬레이터는 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 RPO-RLS보다 개선된 학습 알고리즘을 개발하는 문제 역시 고려하고 있다.

참 고 문 헌

[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*, pp. 1143-1149, 2004, .

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

[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," *Proceedings of the 15th International Conference on Machine Learning*, pp. 278--286, 1998.

[7] V. R. Konda and J. N. Tsitsiklis, *"Actor-Critic Algorithms," SIAM Journal on Control and Optimization*, vol. 42, pp. 1143-1166, 2003.

[8] 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.

저 자 소 개

김종호(Jongho Kim)

2004년 : 고려대학교 제어계측공학과 졸업 (학사)
2004년 ~ 현재 : 고려대학교 제어계측 공학과 대학원

관심분야 : 강화학습, SVM응용
Phone : 019-601-3420
E-mail : oyeasw@korea.ac.kr

강대성(Daesung Kang)

2004년 : 고려대학교 제어계측공학과 졸업 (학사)
2005년 ~ 현재 : 고려대학교 제어계측 공학과 대학원

관심분야 : SVM, 강화학습
phone : 019-506-2086
E-mail : mpkds@korea.ac.kr

박주영(Jooyoung Park)

1983년 : 서울대학교 전기공학과 졸업 (학사)
1985년 : 한국과학기술원 졸업(석사)
1985년 3월~1988년 7월 : 한국전력 월성 원자력발전소 근무
1992년 : University of Texas at Austin 전기 및 컴퓨터공학과 졸업(박사)
1992년8월~1993년2월 : 한국전력 전력 경제연구실 선임전문원
1993년3월~현재 : 고려대학교 과학기술대학 제어계측 공학과 교수

관심분야 : 신경망이론, 지능시스템, 비선형시스템
E-mail : parkj@korea.ac.kr