Using Reinforcement Learning to solve the pendulum swing up

Bruno Costa

July 2, 2014

1 Objective

We want to solve the pendulum swing-up problem [Figure 1] using Reinforcement Learning techiniques, more specifically, using Actor-Critic framework. For that, we will consider both action and state space as continuos, and for such, we will need to use a Function Approximator (FA) for them both. The FA we will use is the tile coding. For the sake of organization, we will split our goal in two:

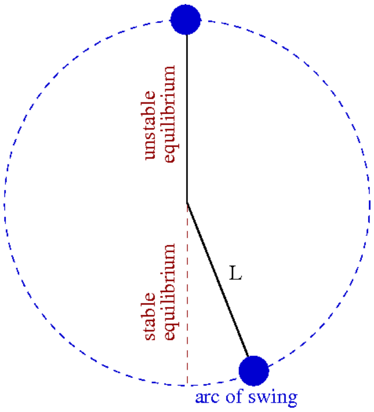

Figure 1: The Pendulum Swing-up problem

Solve the pendulum balancing
In this scenario, the pendulum start on top, and our policy must only balancing it there.

Swing-up
Finally, in this scenario, we will consider the full problem, starting the pendulum on the bottom.

Actor	
α	0.005
Critic	
α	0.1
Parameter	
γ	0.97
λ	0.67
σ	1.0

Table 1: Parameters used in Pendulum Balancing

2 Pendulum Balancing

For the balancing, we changed the code and created a new environment, in which the first observation is $[\pi, 0]$. Using the attributes listed in Table 1 we could make it as we can see in Figures 2, 3 and 4.

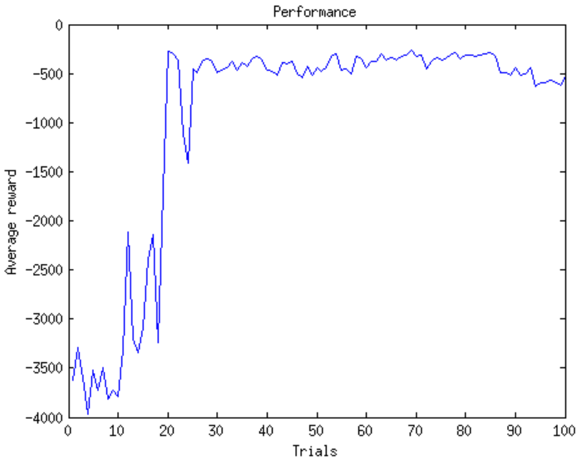

Figure 2: Total reward for each trial in the balacing problem

As we can see, it works(Figure 5). In Figure 2 we can see that the average reward is good after 30 trials, Another interesting thing to notice is the policy in Figure 3: as the angle goes to the left, the action is to push to the right and vice-versa. The critic, in Figure 4 shows that the central position is the best.

3 Pendulum Swing-up

For the swing-up, we used the environment as it is. The first observation is $[0, 0]$. Using the attributes listed in Table **??** (the same attributes here used in the balancing) we could make it as we can see in Figures **??**, **??** and **??**.

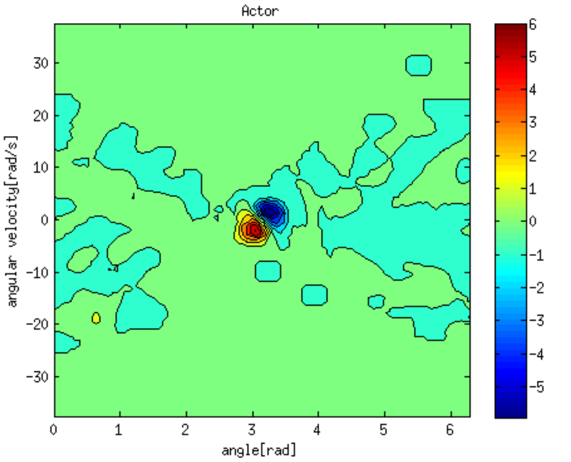

Figure 3: The Actor for the balancing problem

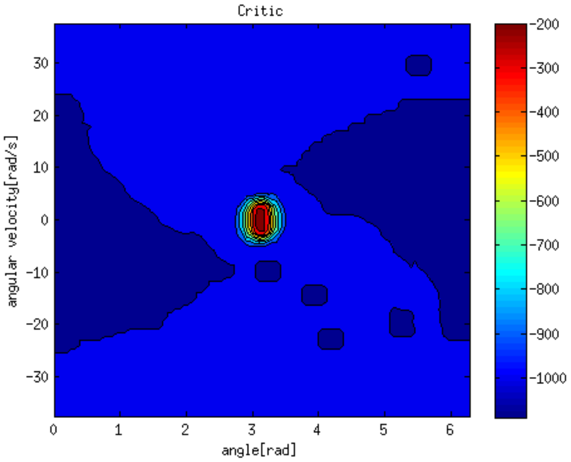

Figure 4: The Critic for the balancing problem

As we can see, it also works(Figure **??**). In Figure **??** we can see that the average reward is good after 130 trials, which is much more than the balancing problem. Another interesting thing to notice is the policy in Figure **??**: there is the swinging, and then, once the pole is on top, it has to balance. The critic, in Figure **??** shows that the central position is the best.

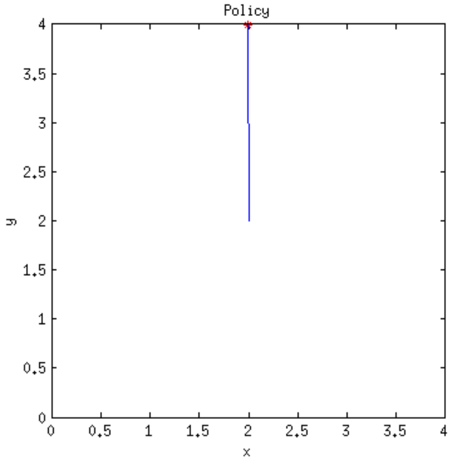

Figure 5: The final position for the balancing problem

Actor	
α	0.005
Critic	
α	0.1
Parameter	
γ	0.97
λ	0.67
σ	1.0

Table 2: Parameters used in Pendulum Swing-up

4 Difficulties

My first version was updating the wrong state, and because of that, it wasn't converging. For each transition, we have the old state (s), the action we took (a), the new state (s') and the reward we get (r). We must update s not s'.

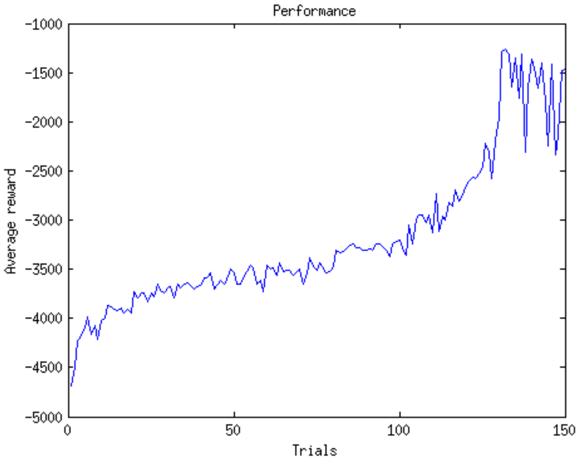

Figure 6: Total reward for each trial in the swing-up problem

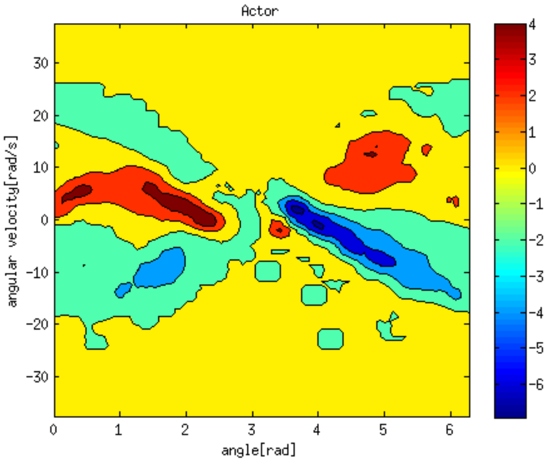

Figure 7: The Actor for the swing problem

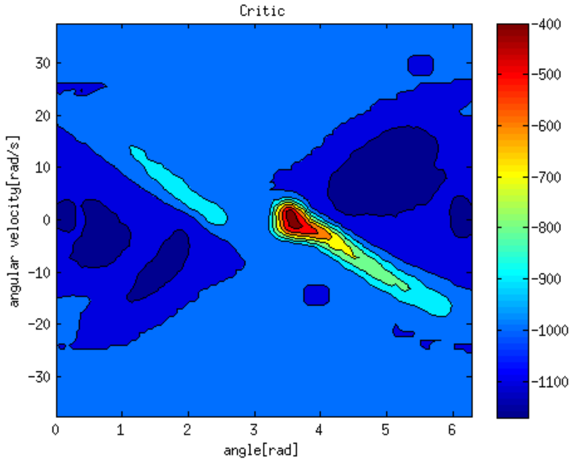

Figure 8: The Critic for the swing problem

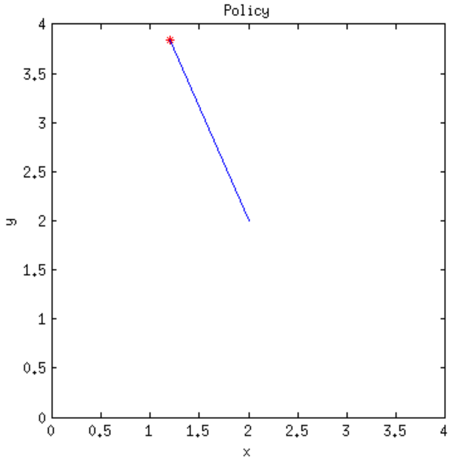

Figure 9: The final position for the swing problem