Programming Assignment 2: Direct Policy Search (DPS)

Amin, Majdi
amin-majdi@student.uml.edu

UMass Lowell, Computer Science — April 1, 2021

1 The model

In the inverted pendulum, all episode terminates ether by being in the vicinity of the goal for 5 consecutive time steps or after 200 time steps.

2 The model

The model for inverted pendulum is this second order ODE:

$$\ddot{\theta} = (u(t) - b\dot{\theta} - mgl\sin(\theta))/2)/(ml^{2/3});$$

And the parameters are:

$$\begin{aligned} m &= 1 \\ l &= 1 \\ g &= 9.82 \\ b &= 0.01 \end{aligned}$$

The boundaries are:

$$\begin{aligned} 0 \leq \theta &< 2\pi \\ -2\pi \leq \dot{\theta} &< 2\pi \end{aligned}$$

Also,The model for cart-pole is made of two second order ODE:

$$\begin{aligned} \ddot{x} &= (2ml\dot{\theta}^2\sin(\theta) + 3mg\sin(\theta)\cos(\theta) + 4f(t) - 4b\dot{x})/(4(M+m) - 3m\cos(\theta)^2); \\ \ddot{\theta} &= (-3ml\dot{\theta}^2 sin(\theta)\cos(\theta) - 6(M+m)g\sin(\theta) - 6(f(t) - b\dot{x})\cos(\theta))/(4l(m+M) - 3ml\cos(\theta)^2) \end{aligned}$$

And the parameters are:

$$\begin{aligned} m &= 0.5 \\ m &= 0.5 \\ l &= 0.5 \\ g &= 9.82 \\ b &= 0.1 \end{aligned}$$

The boundaries are:

$$
\begin{aligned}
0 \leq \theta < 2\pi \\
-2\pi \leq \dot{\theta} < 2\pi \\
-50 \leq x < 50 \\
-5 \leq \dot{x} < 5
\end{aligned}
$$

3 The reward function

According to the article, exponential cost function been used which penalizes the distance to goal:

$$
\begin{aligned}
J(\theta) = E[\textstyle\sum_{t=1}^{T} r(x_t)|\theta] \\
r(x) = \exp\left(-\tfrac{1}{2\sigma_c^2}(x - x_*)^T Q(x - x_*)\right)
\end{aligned}
$$

The weight matrix Q is picked the way that $\dot{\theta}$ and $\dot{x}$ be eliminated from cost functions.

4 The policy representation

To produce a policy parameterization, a normal distribution been used. Both mean and standard deviation of this normal distribution made from a linear function which is theta parameter times features. To produce features, order 3 Fourier bases has been used.

$$
\begin{aligned}
\pi(a|s,\theta) = \tfrac{1}{\sigma(s,\theta)\sqrt{2\pi}} \exp\left(-\tfrac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right) \\
\mu(s,\theta) = \theta_\mu^T x_\mu(s) \\
\sigma(s,\theta) = \exp\left(\theta_\sigma^T x_\sigma(s)\right)
\end{aligned}
$$

5 The optimization algorithm

In this work, CMA-ES black-box algorithm has been used. The initial sigma is equal to 0.25 for both pendulum and cart-pole.

6 Plots for inverted pendulum

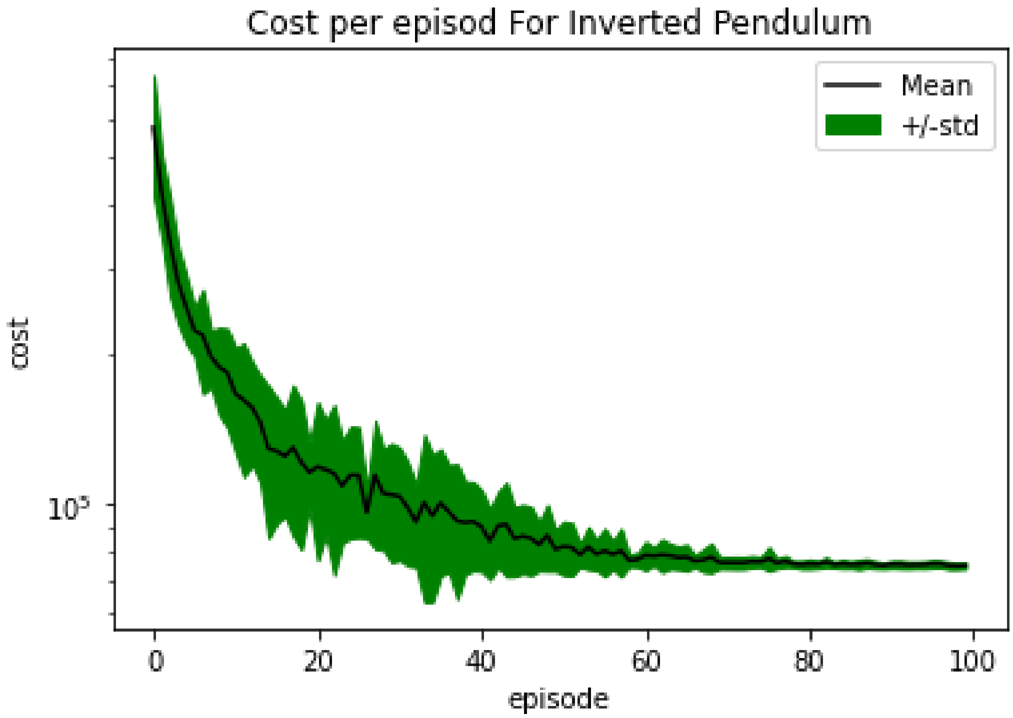

Fig.1

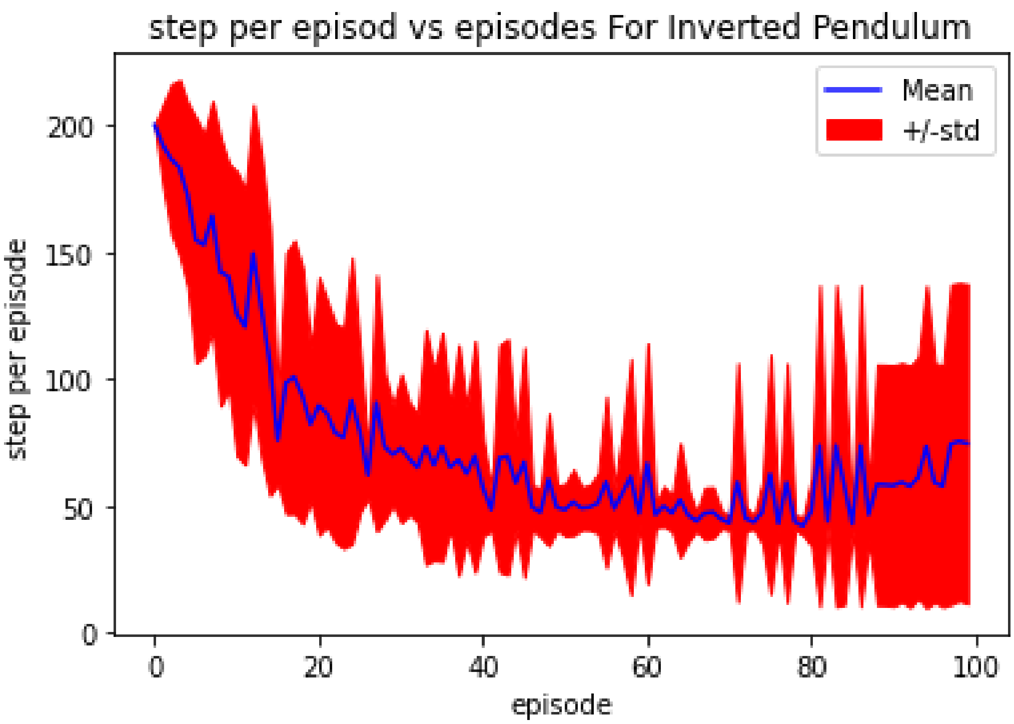

Fig.2

7 Plots for cart-pole

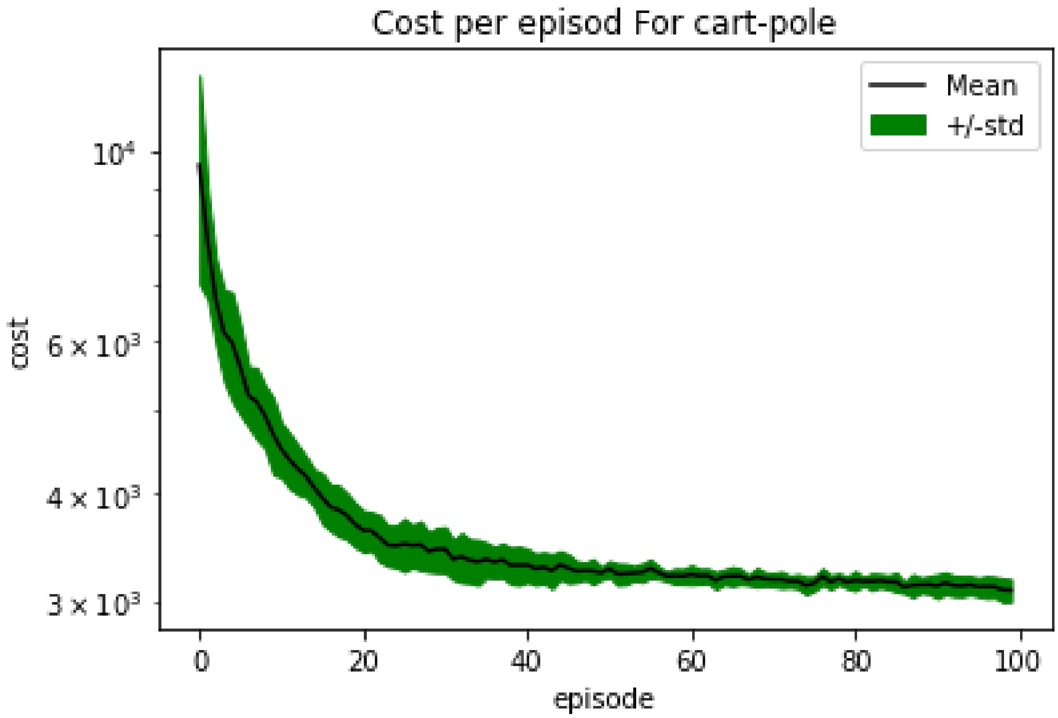

Fig.3

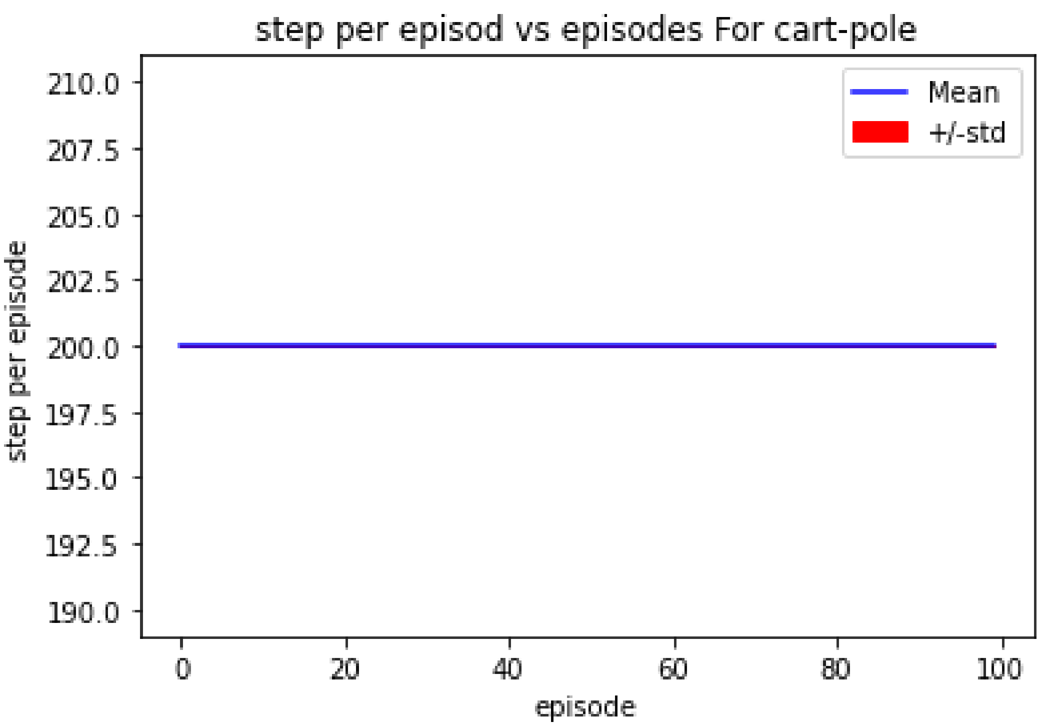

Fig.4

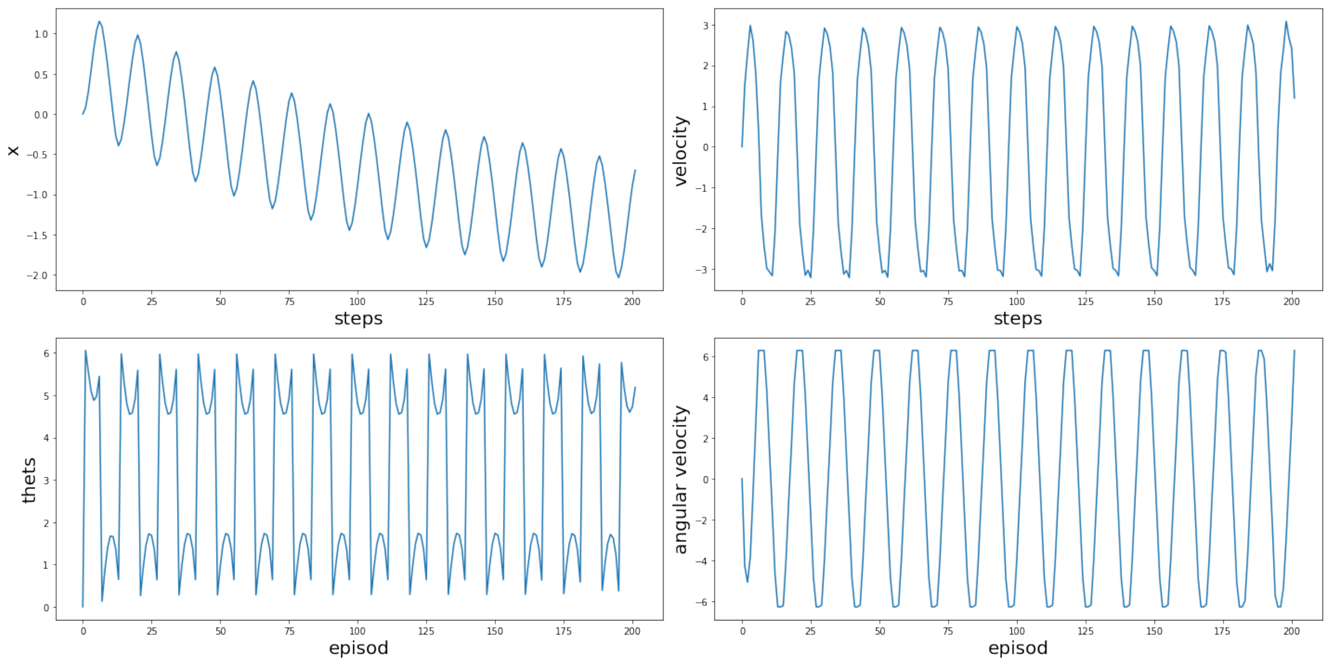

Fig.5

8 Comparison

8.1 Inverted pendulum

Figure 1 shows that the cost is decreasing and after 60 episode it is firmly fixed. In this configuration,as the episodes terminate after 200 steps,the decrease of mean in fig 2 shows that our agent can reach to the goal even after 1 episode and we can confidently say that after 20 episodes all agents see the goal. The figure 2 in the article shows the best result so far. So our best result is very good comparable to that.

8.2 Cart-pole

as we can see in Figure 3 the cost is minimized by the optimizer and from Figure 5 it is obvious that the agent is learning a policy, but this is not the optimal policy. Both x and θ are swinging around their goal position, so the failure in finding optimal policy is either due to a weak cost function or small step size in each episode. If we give the algorithm more time to learn, it may learn the optimal policy. The run-time for learning process of cart-pole with this configuration took 4 hours. Therefore, the policy representation with normal distribution and Fourier basis features is not time efficient.