TP3 for Reinforcement Learning

Leman FENG
Email: flm8620@gmail.com
Website: lemanfeng.com

December 17, 2017

Q1 & Q2

I choose by default $N = 100, T = 100$, and initialize with $\theta = -0.1$.

For the constant step gradient descent, step = 0.001 diverges, step=0.0001 converges(Figure 1).

For the Annealing Step. I defined the schema as step $= \frac{\alpha}{t+10}$ where α is the learning rate and t is the iteration. Compare to Constant step schema, Annealing step can tolerate a large learning rate, but it converges slowly (Figure 2). We can see with $N = 100, \alpha = 0.01$, the performance is good.

For Adam Step, this algorithm is more robust. I tried with $\alpha = 0.01$ and $N = 100, n_iter = 100$, and again with $N = 10, n_iter = 1000$. Both tests converged (Figure 3). $N = 10$ will lead to a high variance of gradient. With same step = 0.01, Annealing stepper will diverge so fast, but Adam stepper is still stable. So we can say Adam stepper is suitable for stochastic gradient method.

To answer to Question 1, small N will lead to large variance, so small N should be used with small step α. As shown in Figure 2, with $N = 100, \alpha = 0.01$, annealing stepper converges. But then I changed N to 20, it diverged. Again with a small step $\alpha = 0.0005$, Annealing stepper can converge with $N = 20$

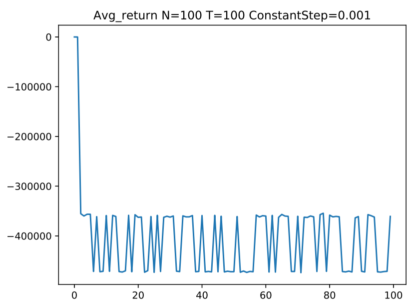
Q3

To have some generality, I changed ϕ to $[1, s, a, s^2, a^2, sa]$. FQI still works very well. It converges to theoretical optimum in 5 iterations(Figure 4). $J(\pi_k)$ is shown in Figure 5

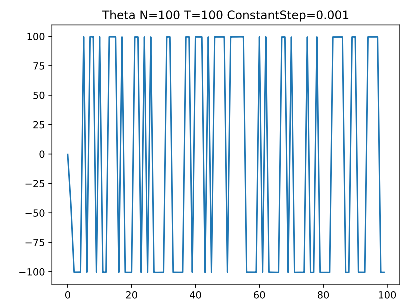
Cart-pole

I found the cart-pole code, so I applied Policy gradient method to cart-pole. Action space is $\{0, 1\}$, state space is $\mathbb{R}^4$. I use Gibbs model and set

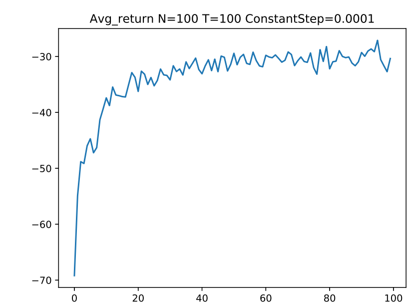
$$Q_\theta(s, a = 0) = \theta_0^T s, Q_\theta(s, a = 1) = \theta_1^T s,$$



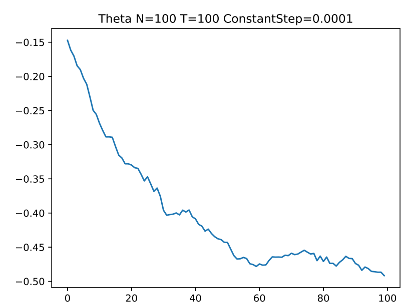
(a) Avg Return, Constant Step 0.001



(b) Theta, Constant Step 0.001

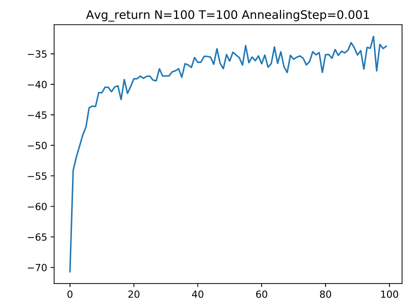


(c) Avg Return, Constant Step 0.0001

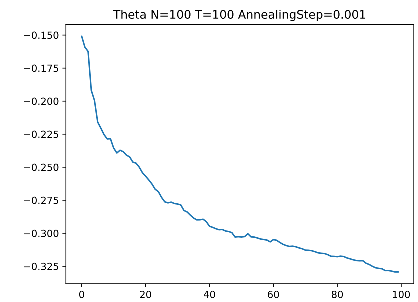


(d) Theta, Constant Step 0.0001

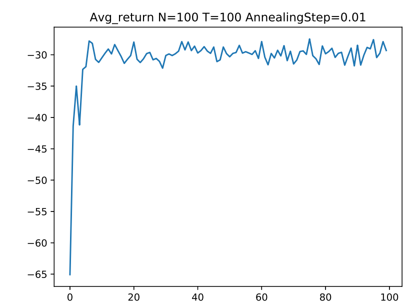
Figure 1: Constant Step



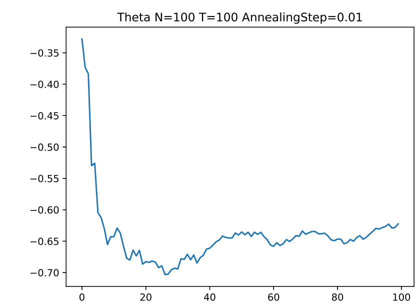
(a) Avg Return, Annealing Step 0.001



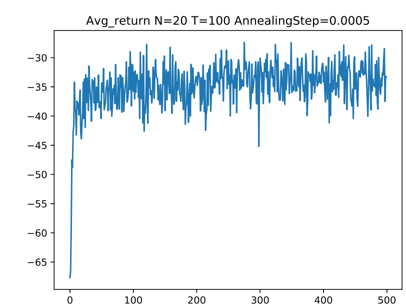
(b) Theta, Annealing Step 0.001



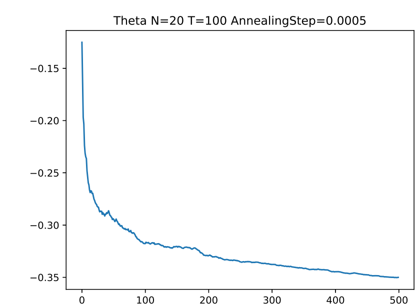
(c) Avg Return, Annealing Step 0.01



(d) Theta, Annealing Step 0.01

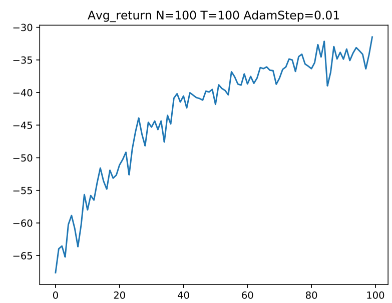


(e) Avg Return, Annealing Step 0.0005

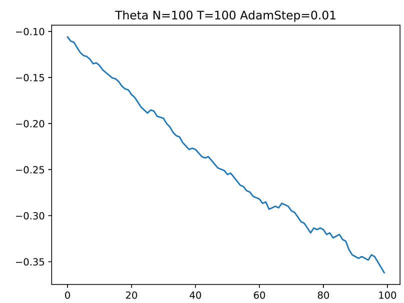


(f) Theta, Annealing Step 0.0005

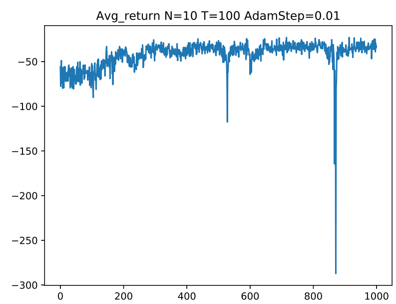
Figure 2: Annealing Step



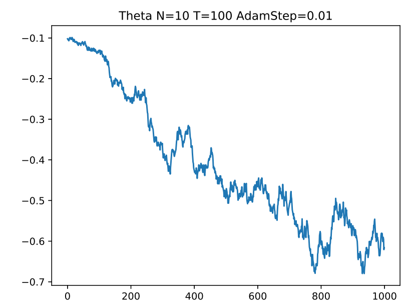
(a) Avg Return, Adam Step N=100



(b) Theta, Adam Step N=100

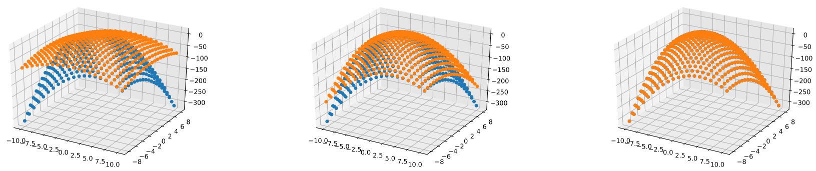


(c) Avg Return, Adam Step N=10



(d) Theta, Adam Step N=10

Figure 3: Adam Step



(a) Q learnt at iteration 1 (b) Q learnt at iteration 2 (c) Q learnt at iteration 5

Figure 4: Q learnt

Because we want cart-pole to stay verticle as long as possible, so I set $\gamma = 1$. I use Adam stepper with $N = 30, \alpha = 0.1$. Result is shown in Figure 6

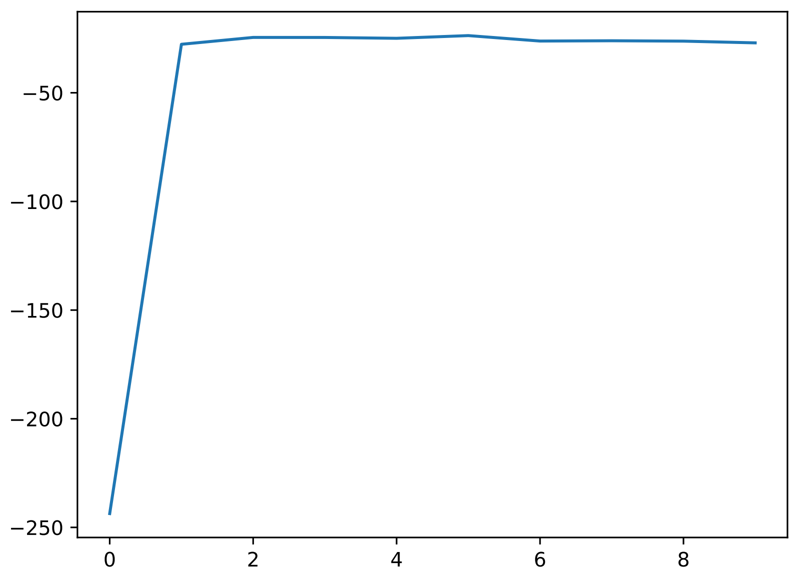


Figure 5: $J(\pi_k)$

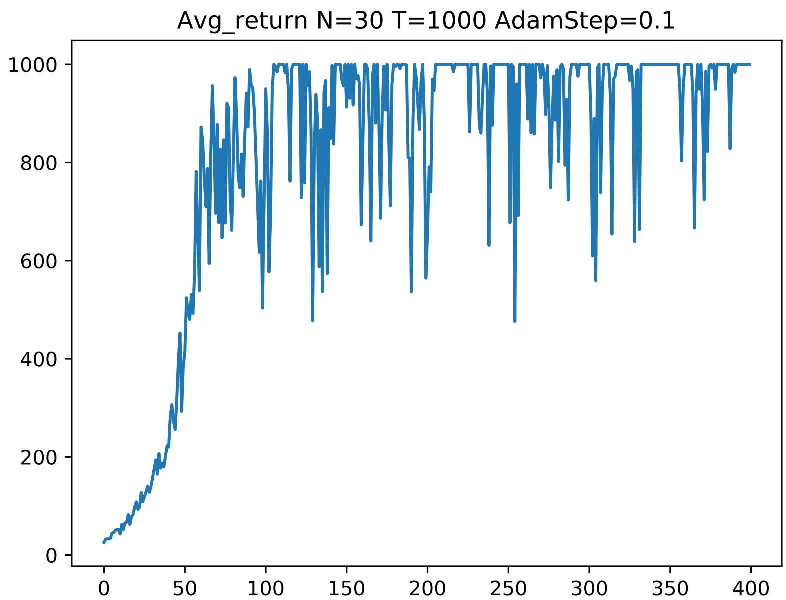


Figure 6: cart-pole