Variance Reduction for Policy Gradient with Action-Dependent Factorized Baselines

Topics in Machine Learning
Course Project
Team Pheonix

25 November 2019

1. Problem Statement

Policy gradient methods have enjoyed great success in deep reinforcement learning but suffer from high variance of gradient estimates. The high variance problem is seen particularly in problems with long horizons or high-dimensional action spaces. To mitigate this issue, we derive a bias-free action-dependent baseline for variance reduction which fully exploits the structural form of the stochastic policy itself and does not make any additional assumptions about the MDP. Through this project we aim to demonstrate and quantify the benefit of the action-dependent baseline through both theoretical analysis as well as numerical results, including an analysis of the suboptimality of the optimal state-dependent baseline.

2. Our Environment

Introduction:
We have taken Bipedal Walker , which an Open Gym Environment, which basically consists of two legs each with a joint.Our goal is to teach the Bipedal-walker to walk by applying the torque on these joints. Therefore the size of our action space is 4 which is torque applied on 4 joints. You can apply the torque in the range of (-1, 1).

Reward Structure : The agent gets a positive reward proportional to the distance walked on the terrain. It can get a total of + 300 reward all the way up to the end. If agent tumbles, it gets a reward of -100. There is some negative reward proportional to the torque applied on the joint,this is to ensure that agent learns to walk smoothly with minimal torque.

3. Algorithm

In the following, we analyze action-dependent baselines for policies with conditionally independent factors. For example, multivariate Gaussian policies with a diagonal covariance structure are commonly used in continuous control tasks. Assuming an m-dimensional action space, we have :

In this case, we can set bi, the baseline for the ith factor, to depend on all other actions in addition to the state. Let at(i) denote all dimensions other than i in at and denote the ith baseline by bi(st , a(t)i) : Due to conditional independence and the score function estimator, we have:

Now, given we can finally write our function as :

This is compatible with advantage function form of the policy gradient . Takes the form of Actor Critic Method:

Finally on solving for the optimal baseline case ,

we get the following the part :

Now, combining all of the above results we have our final baseline implementation :

4. Experiment 1 : Only with baseline(b(st)) Actor Critic

We first tried implementing the Bi-Pedal system with normal baseline b(s(t) , we obtained the following results :

5. Experiment 2 : Action Baseline(b(st,at)

Given the four actions, we sub-sample each ith action and while calculating we calculate the average of actions for ith part and replace it in the average episode part .