Proximal Policy Optimization with Dynamic Clipping

Student: Rishikesh Vaishnav
Mentor: Sicun Gao

August 11, 2018

Introduction

Reinforcement Learning

- A general algorithmic technique that seeks to replicate behavioral learning.
- Attempts to maximize rewards through episodic sequences of actions.

Introduction (contd.)

Trust Region Policy Optimization

- TODO explain TRPO's connection to Reinforcement Learning
- The theory behind TRPO suggests choosing a policy parameterization θ maximizing the surrogate loss:

$$L_{\theta_{old}}(\theta) - CD_{KL}^{max}(\theta, \theta_{old})$$

where C is a fixed positive constant and it is shown that

$$L_{\theta_{old}}(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim p_{\theta_{old}}, a\sim\theta_{old}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)}A_{\theta_{old}}(s,a)\right]$$

where $p_{\theta_{old}}$ is the normalized discounted visitation frequency distribution.

- In theory, doing so guarantees monotonic improvement of the policy.

Introduction (contd.)

Proximal Policy Optimization

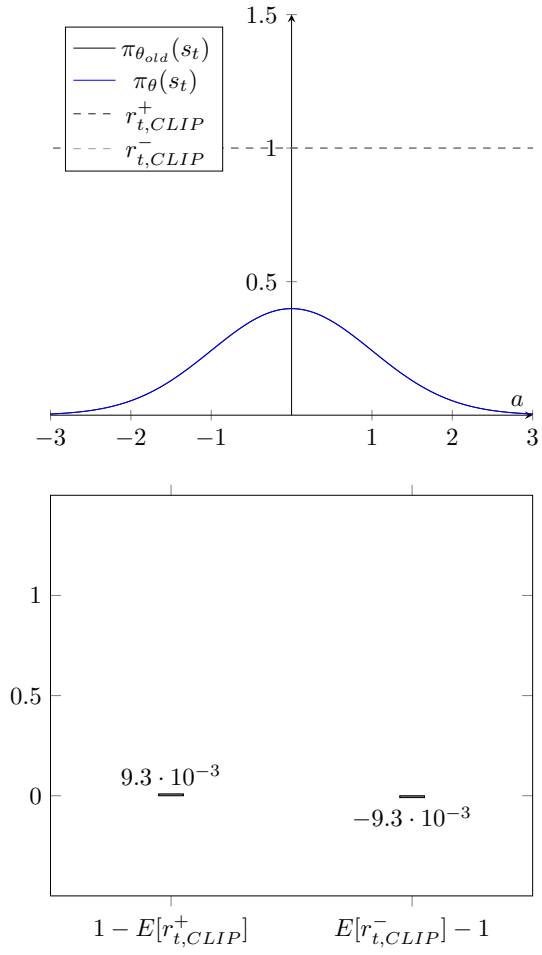
- TODO explain connection of PPO to TRPO

Potential Shortcomings of PPO

- We can keep track of the expected loss contributions from positive and negative advantages as we get further from the mean.
- TODO explain expected loss contribution equations
- The major effect of using a clipper is to increase expected loss contributions as we get further from the mean.
- Any min-filter that accomplishes this should be valid.

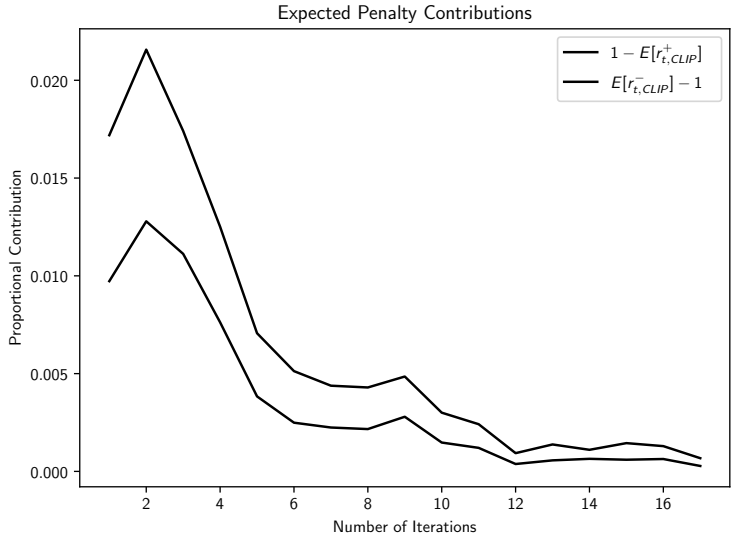
Potential Shortcomings of PPO (contd.)

What happens as we learn?



Potential Shortcomings of PPO (contd.)

- Clearly, there is a growing discrepency between expected loss contributions from positive and negative estimators as we move farther from the mean.
- This discrepancy exists empirically as well:



- TODO explain why this does not manifest itself in the actual loss.
- However, this discrepancy is not inherent to the TRPO surrogate loss. We can imagine that losses are distributed approximately equally.

Idea

- Is there a way to effectively control this expected discrepancy, along with the rate at which the expected proportional penalty increases?
- TODO introduce idea

Results

TBA