Proximal Policy Optimization with Dynamic Clipping

Student: Rishikesh Vaishnav
University of California, San Diego

Mentor: Sicun Gao, Ph.D
University of California, San Diego

Abstract

Proximal Policy Optimization (PPO), a policy gradient algorithm drawing closely from the theory supporting Trust Region Policy Optimization (TRPO), has emerged as one of the most effective tools in reinforcement learning (RL) problems. PPO makes use of a loss function with a clipped importance sampling ratio using a single parameter ϵ. Although PPO shows promising empirical performance, it is vulnerable to problem-specific imbalances in its handling of positive and negative advantages. We investigate one such imbalance, addressing a discrepancy in expected penalty contributions of positive and negative estimators. By precisely calculating this discrepancy and minimizing it before each iteration, we empirically demonstrate that eliminating this discrepancy can improve the overall performance of PPO.

1 Introduction

Note: This paper assumes knowledge of the common terms and concepts in reinforcement learning (see [TODO cite sutton and barto]).

Among current reinforcement learning (RL) algorithms, Policy Gradient methods have seen significant success at a wide range of tasks. These methods seek to learn performant policy distributions directly, rather than learning a value function to indirectly guide the policy [TODO cite paper covering policy gradient]. Using a policy $\pi_\theta(a)$ parameterized by θ, these algorithms have the general form:

Algorithm 1 Generic Policy Gradient

```
Initialize $\theta$ arbitrarily
while True do                                  ▷ loop forever
    $\theta_{old} \leftarrow \theta$
    $rollout \leftarrow (s, a, r)$ from multiple
        episodes following $\pi_\theta$
    Set $\theta$ to maximize the loss function
        $L(rollout, \theta, \theta_{old})$
end while
```

Within this framework, the choice of a loss function is the key to an effective algorithm, and current research in reinforcement learning focuses particularly on finding new ways to represent this loss. A recent innovation in policy gradient methods is Trust Region Policy Optimization (TRPO), which approximates the updates perfomed by an agent that uses the following loss function to guarantee monotonic improvement in the policy's performance:

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

, where C is a constant,

$$L_{\theta_{old}}(\theta) = \mathbb{E}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} A_t \right]$$

, and A_t is the advantage at time t [TODO cite]. In TRPO implementations, this loss function results in relatively small step sizes, so in practice the penalty term $CD_{KL}^{max}(\theta, \theta_{old})$

More fascinating text. Features[1] galore, plethora of promises.

2 This is Another Section

Some embedded literal typset code might look like the following :

```
#include <iostream>
using namespace std;
main()
{
cout << "Hello world \n";
return 0;
}
```

Now we're going to cite somebody. Watch for the cite tag. Here it comes [1].

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Notes

[1] Remember to use endnotes, not footnotes!

References

[1] `https://arxiv.org/abs/1707.06347`

[2] `https://arxiv.org/abs/1506.02438`

[3] `https://arxiv.org/abs/1502.05477`