MASTER THESIS

Exploring the impact of markets on the credit assignment problem in a multiagent environment

Zarah Zahreddin

Entwurf vom August 22, 2021

MASTER THESIS

Exploring the impact of markets on the credit assignment problem in a multiagent environment

Zarah Zahreddin

Professor:	Prof. Dr. Claudia Linnhoff-Popien
Supervisor:	Kyrill Schmid
	Robert Müller
Submission Date:	31. December 2021

I hereby affirm that I wrote this Master Thesis on my own and I did not use any other sources and aids than those stated.

Munich, 31. December 2021

. . . .
Signature

Abstract

Hier kommt der Abstract hin. Hier kommt der Abstract hin.

Contents

1 Introduction

- Motivation
- Goal
- Research Question
- Structure

2 Related Work

- Definiton of field of research
- Scientific Scope
- Which comparable work in research exists?
- Seperation from other works

2.1 Reinfocement learning

2.1.1 Environment Interaction

Describe and explain the reward system -> Agent observes, takes action, env reacts and returns reward

2.1.2 Policy Optimization vs Action selection optimization

Describe and explain the reward system -> Agent observes, takes action, env reacts and returns reward

2.1.3 Multiagent settings

Describe and explain the settings, mixed/cooperation

2.2 CAP

2.3 Markets

3 Background

- Describe the technical basis of your work
- Do not tell a historical story - make it short

3.1 PPO

In 2017 Schulman et al. introduced the concept of Proximal Policy Optimization (PPO) in "Proximal Policy Optimization Algorithms"[SWD$^+$17] and this section is solely based on that article. Policy optimization is the improvement of the action selection strategy π based on the current state s_t. This is achieved by rotating two steps: 1. sampling data from the policy and 2. optimizing that data through several epochs.

The origin of PPO lies in a similar approach called Trust Region Policy Optimization (TRPO). TRPO strives to maximize the following function:

$$maximize_{\theta} \hat{\mathbb{E}}_t[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}\hat{A}_t - \beta \, KL[\pi_{\theta_{old}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]] \tag{3.1}$$

with $\hat{A}_t$ as an estimator of the Advantage function. The fraction in the Minuend can be replaced by $r(\theta)$ and represents the probability ratio of an action in the current policy in comparison to the old policy. The result of $r(\theta)$ is greater than one, if an action is very probable in the current policy. Otherwise the outcome lies between zero and one. Schulman et al. further describe that TRPO maximizes the "surrogate" objective

$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}\hat{A}_t] = \hat{\mathbb{E}}_t[r(\theta)\hat{A}_t] \tag{3.2}$$

which, maximized on its own without a penalty, results in a large outcome and leads to a drastic policy update.

In order to stay in a trust region, as the name suggests, a penalty is subtracted from the surrogate function (3.2). The penalty is the Subtrahend of equation (3.1) and contains the fixed coefficient β. Regardless of the function details and outcome of KL, the coefficient β is hard to choose, since different problems require different penalty degrees. Even in a single problem it could be necessary to adapt the coefficient, due to changes within the setting.

Therefore Schulman et al. introduced

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[\min(r(\theta)\hat{A}_t, clip(r(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \tag{3.3}$$

which is very similar to (3.1) but does not require coefficients. The first part of min

contains L^{CPI} (3.2). The second part contains a *clip* function which narrows the space of policy mutation with the small hyperparameter ϵ. After applying the clip function $r(\theta)$ lies between $[1 - \epsilon, 1 + \epsilon]$. Calculating the minimum of the clipped and unclipped probability ratio produces the lower bound of the unclipped $r(\theta)$, preventing the policy to change too much. [TODO ???]

PPO is defined by the following equation

$$L_t^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)] \tag{3.4}$$

PPO Algo

with c_1 and c_2 as coefficients. Schulman et al. point out that the loss function $L_t^{VF} = (V_\theta(s_t) - V_t^{targ})^2$ combines the policy surrogate and the value function error term and is necessary once a neural network shares parameters between policy and value function. Finally an entropy bonus S is added to ensure exploration. Schulman et al. continues to show an example of an Algorithm using PPO, see Fig. 3.1. N detonates (parallel) actors collecting data in T timesteps in each Iteration. Afterwards the policy is optimized in K epochs by computing the Loss function (3.4) on the corresponding NT timesteps of data, using a minibatch.

Algorithm 1 PPO, Actor-Critic Style

---
   for iteration=$1, 2, \dots$ **do**
      for actor=$1, 2, \dots, N$ **do**
         Run policy $\pi_{\theta_{old}}$ in environment for T timesteps
         Compute advantage estimates $\hat{A}_1, \dots, \hat{A}_T$
      end for
      Optimize surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
      $\theta_{old} \leftarrow \theta$
   end for

Figure 3.1: Exemplary use of PPO, as shown in "Proximal Policy Optimization Algorithms"[SWD+17]

3.2 DQN

4 Concept

- What is your plan?
- How do you proof that it worked? -> Metric and Experiments

5 Implementation

- How exactly did you do it?
- Experiment parameters
- Experiment setup
- No need to mention framework, software libraries or tools

6 Results

- Result presentation
- Description of images and charts

7 Discussion

- Are the findings as expected?
- Why are the things as they were observed?
- New experiments that provide further insights
- Make your results more comprehensible

8 Conclusion

(Briefly summarize your work, its implications and outline future work)
- What have you done?
- How did you do it?
- What were the results?
- What does that imply?
- Future work

List of Figures

List of Tables

Listings

Bibliography

[SWD+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.