Theoretical Foundations of A3C and TRPO in Reinforcement Learning

Manuel Mateo Delgado-Gambino Lopez
Advanced Automatic Learning
UIE
Email: manuel.delgado_gambino.01@uie.edu

***Abstract*—This paper systematically reviews the mathematical foundations of two advanced reinforcement learning algorithms: Asynchronous Advantage Actor-Critic (A3C) and Trust Region Policy Optimization (TRPO). We present formal derivations of their key equations, including policy gradient updates and trust region constraints, while emphasizing their theoretical guarantees and practical implementations. Formatted in IEEE LaTeX style, this work serves as a concise reference for researchers and practitioners implementing modern policy optimization techniques.**

I. Introduction

Modern reinforcement learning (RL) has witnessed significant advancements through policy gradient methods that balance exploration and convergence guarantees. Two pivotal algorithms in this domain are:

- **A3C (Asynchronous Advantage Actor-Critic)**: A distributed framework leveraging parallel actors to decorrelate samples and accelerate learning through asynchronous updates [**?**]. Combines value function estimation with policy optimization through advantage-weighted updates.
- **TRPO (Trust Region Policy Optimization)**: A theoretically grounded approach ensuring monotonic policy improvement via constrained optimization over a trust region [**?**]. Addresses the challenge of destructive policy updates through KL-divergence constraints.

This work contributes a unified mathematical presentation of these algorithms' core components, facilitating comparative analysis and implementation.

II. Asynchronous Advantage Actor-Critic (A3C)

A. Architectural Overview

A3C employs multiple parallel agents interacting with environment instances, asynchronously updating a global network. This parallelism achieves data efficiency while maintaining diversity in experience sampling.

B. n-Step Advantage Estimation

The advantage function estimates the relative value of actions using temporal difference learning over k steps:

$$A(s_t, a_t; \theta, \theta_v) = \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}; \theta_v) - V(s_t; \theta_v) \quad (1)$$

where $\gamma \in [0, 1]$ is the discount factor, $V(s; \theta_v)$ denotes the value function parameterized by θ_v, and k governs the bias-variance tradeoff.

C. Policy Gradient Optimization

The policy parameters θ are updated using advantage-weighted gradient ascent:

$$\nabla_\theta \mathcal{L}_{\text{policy}} = \mathbb{E}\left[\nabla_\theta \log \pi(a_t|s_t; \theta) A(s_t, a_t; \theta, \theta_v)\right] \quad (2)$$

D. Entropy Regularization

To prevent premature convergence, entropy regularization encourages exploration:

$$\mathcal{L}_{\text{entropy}} = \beta \mathbb{E}\left[H(\pi(\cdot|s_t; \theta))\right] \quad (3)$$

where β controls regularization strength and entropy H is:

$$H(\pi) = -\sum_{a \in \mathcal{A}} \pi(a|s_t; \theta) \log \pi(a|s_t; \theta) \quad (4)$$

III. Trust Region Policy Optimization (TRPO)

A. Constrained Policy Improvement

TRPO maximizes a surrogate objective $\eta(\pi)$ while constraining policy divergence:

$$\max_\pi \mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}, a \sim \pi_{\text{old}}} \left[\frac{\pi(a|s)}{\pi_{\text{old}}(a|s)} A_{\pi_{\text{old}}}(s, a)\right] \quad (5)$$

subject to:

$$\mathbb{E}_{s \sim \rho_{\pi_{\text{old}}}}\left[D_{KL}(\pi_{\text{old}}(\cdot|s) \parallel \pi(\cdot|s))\right] \leq \delta \quad (6)$$

where δ is the trust region radius.

B. Natural Policy Gradient

TRPO approximates the natural gradient using conjugate gradient descent with Fisher information matrix F:

$$\theta_{k+1} = \theta_k + \alpha F^{-1}(\theta_k)\nabla_\theta \eta(\pi) \tag{7}$$

where $\alpha = \sqrt{\frac{2\delta}{\nabla_\theta \eta(\pi)^T F^{-1} \nabla_\theta \eta(\pi)}}$ adapts the step size.

IV. Comparative Analysis

A3C provides practical advantages through distributed sampling and empirical stability, while TRPO offers strong convergence guarantees via constrained optimization. The former excels in diverse environments requiring exploration, whereas the latter is preferred in safety-critical applications needing update stability.

V. Conclusion

This work formalizes the mathematical underpinnings of A3C and TRPO, two cornerstone algorithms in modern reinforcement learning. By presenting their objective functions, constraints, and update rules within a unified framework, we enable systematic comparison and informed algorithm selection. Future extensions could incorporate proximal policy optimization (PPO) and distributed TRPO variants.