Journal of Computational Mathematics
Vol.40, No.6, 2022, 1006–1032.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2104-m2021-0007

A STOCHASTIC TRUST-REGION FRAMEWORK FOR POLICY OPTIMIZATION*

Mingming Zhao, Yongfeng Li and Zaiwen Wen[1)]
Beijing International Center for Mathematical Research, Peking University, China
Email: mmz102@pku.edu.cn, yongfengli@pku.edu.cn, wenzw@pku.edu.cn

Abstract

In this paper, we study a few challenging theoretical and numerical issues on the well known trust region policy optimization for deep reinforcement learning. The goal is to find a policy that maximizes the total expected reward when the agent acts according to the policy. The trust region subproblem is constructed with a surrogate function coherent to the total expected reward and a general distance constraint around the latest policy. We solve the subproblem using a preconditioned stochastic gradient method with a line search scheme to ensure that each step promotes the model function and stays in the trust region. To overcome the bias caused by sampling to the function estimations under the random settings, we add the empirical standard deviation of the total expected reward to the predicted increase in a ratio in order to update the trust region radius and decide whether the trial point is accepted. Moreover, for a Gaussian policy which is commonly used for continuous action space, the maximization with respect to the mean and covariance is performed separately to control the entropy loss. Our theoretical analysis shows that the deterministic version of the proposed algorithm tends to generate a monotonic improvement of the total expected reward and the global convergence is guaranteed under moderate assumptions. Comparisons with the state-of-the–art methods demonstrate the effectiveness and robustness of our method over robotic controls and game playings from OpenAI Gym.

Mathematics subject classification: 49L20, 90C15, 90C26, 90C40, 93E20.
Key words: Deep reinforcement learning, Stochastic trust region method, Policy optimization, Global convergence, Entropy control.

1. Introduction

In reinforcement learning, the agent starts from an initial state and interacts with the environment by executing an action from some policy iteratively. At each time step, the environment transforms the current state into the next state with respect to the action selected by the agent and gives back a reward to the agent to evaluate how good the action is, then the agent makes a new action for the next interaction based on the feedback. Repeating the above transition dynamics generates a trajectory where stores the visited states, actions and rewards. During the interactions, the transition probability and the reward function are totally determined by the environment, but the intrinsic mechanism may be mysterious. The policy characterizes the distribution of actions at each possible state. The problem is how to design a policy for the agent to maximize the total expected reward along a trajectory induced by the policy. The state-of-the-art model-free methods for reinforcement learning [28, 33] can be

* Received January 7, 2021 / Accepted April 22, 2021 /
Published online November 16, 2021 /
1) Corresponding author

divided into policy-based and value-based methods. Policy-based methods directly learn or try to approximate the optimal policy by policy improvement and policy evaluation alternatively. They generate a map, i.e., a distribution from states to actions, which can be stochastic or deterministic. That is, they can be applied to both continuous and discrete action spaces. While in value-based methods, the goal is approximating the solution of the optimal Bellman equation based upon the temporal difference learning [33]. They learn a value function defined on the state-action pairs to estimate the maximal expected return of the action taken in the state. Then at each sate, the optimal policy based on the value function predicts a single action by maximizing the values.

The recent progress of deep neural networks [23] provides many scalable and reliable learning based approaches [8, 11, 24, 29, 31] for solving large and complex real-world problems in reinforcement learning. The curse of dimensionality is conquered by expressing the value and/or policy function with a deep neural network from high-dimensional or limited sensory inputs. The deepening expedites the evolution of end-to-end reinforcement learning, also referred as deep reinforcement learning. As a representative and illuminative algorithm in deep value-based methods, deep Q-learning (DQN) [25] has succeeded in many discrete domains such as playing Atari games. The learned agent arrives at a comparable level to that of a professional human games player. They construct a Q-network to receive the raw pictures as inputs, and optimize the weights by minimizing the Bellman residual. DQN can be viewed as a deep value iteration method directly, and some independent improvements including their combinations have been summarized in [14]. The success of DQN and its variants has a restriction on the type of the problem, specifically, the maximal operator in the objective function makes the optimization to be less reliable in continuous and/or large action space. By representing the greedy action selection with a policy network, the deep deterministic policy gradient (DDPG) method [24] successfully extends the algorithmic idea of DQN into the continuous action space. The value network imitates the training in DQN and the policy network is updated by maximizing the estimated values. The two delayed deep deterministic (TD3) policy gradient algorithm [9] substantially improves DDPG by building double deep Q-networks to avoid overestimation in value estimates and delaying the policy updates to reduce the per-update error in DDPG.

Different from the optimization models based on value functions, policy-based algorithms also concentrate on optimizing the policy iteratively. In the policy improvement step, the actor updates the policy by optimizing an appropriate objective function using gradient-based methods. Policy evaluation creates a critic, i.e., a value function, to assess the policy by minimizing the Bellman error associated with the policy, which provides a basis for policy improvement. Thus the policy-based methods are usually classified as actor-critic methods. As the optimization is practically based on the observations, the generalized advantage estimators (GAE) [30] are mostly considered for the bias-variance tradeoff and numerical stability. The discrepancy among the state-of-the-art policy-based methods mainly locates in the actor part, specifically, the surrogate function used for improving the policy. The trust region policy optimization (TRPO) [29] generalizes the proof the policy improvement bound in [16] into general stochastic policies and proposes a trust region model for policy update. The model function is a local approximation of the total expected reward and the Kullback-Leibler (KL) divergence between two policies is considered as a distance constraint. The subproblem under parameterization is highly nonlinear and nonconvex rather than a typical quadratic model as in [3, 6, 37] because the policy is parameterized by a neural network and the trust region constraint is replaced by a distance function of two policies. In order to develop a practical